Proceedings

EIGHTH ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

October 7-9, 1970

University of Illinois
Allerton House, Monticello, Illinois
Sponsored by the
Department of Electrical Engineering and the
Coordinated Science Laboratory of the
University of Illinois, Urbana, Illinois and the
Circuit Theory and Automatic Control Groups of the
Institute of Electrical & Electronics Engineers

IEEE Catalog No. 70 C 42-CT

ALLERTON CONFERENCE ON CIRCUIT AND
SYSTEM THEORY
Eighth, 1970

1 WEEK CHECKOUT

00320

DIGITAL COMPUTER
LABORATORY
LIBRARY

PROCEEDINGS
EIGHTH ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

A. H. Haddad
B. C. Kuo
G. Metze
Conference Co-Chairmen

Conference held
October 7, 8, 9, 1970
Allerton House
Monticello, Illinois

Sponsored by the

DEPARTMENT OF ELECTRICAL ENGINEERING
and the
COORDINATED SCIENCE LABORATORY
of the
UNIVERSITY OF ILLINOIS
Urbana, Illinois 61801

and the

CIRCUIT THEORY AND AUTOMATIC CONTROL GROUPS
OF THE
INSTITUTE OF ELECTRICAL & ELECTRONICS ENGINEERS

[redacted]

DIGITAL COMPUTER
LABORATORY
LIBRARY

FOREWORD

The Eighth Annual Allerton Conference on Circuit and System Theory was held at Allerton House from October 7-9, 1970. These Proceedings are a permanent record of the papers presented at that Conference. The papers described new results in information and coding theory, digital systems, communication sciences, control systems, network theory, and computer-aided design.

In addition to the technical sessions, the remoteness of Allerton Park continues to provide a setting fruitful to the exchange of ideas throughout the Conference.

The Conference was co-sponsored by the Department of Electrical Engineering and the Coordinated Science Laboratory of the University of Illinois, and the IEEE Groups on Circuit Theory and Automatic Control. We acknowledge the encouragement and support of Professor E. C. Jordan, Head of the Department of Electrical Engineering at the University of Illinois. We thank Mr. Harold B. Lawler for his help with certain business aspects of the Conference. We are grateful to secretaries, Mrs. Rosa Townsend, Rose Washburn, Rose Lane, and Sherry Kallembach for their assistance. Also, we thank P. V. Kokotovic for organizing the special session on singular perturbation. In addition, we are indebted to our colleagues for their vital assistance: R. T. Chien, W. Mayeda, W. R. Perkins, F. Preparata, J. B. Cruz, Jr., and N. Wax.

A. H. Haddad
B. C. Kuo
G. Metze

TABLE OF CONTENTS

Foreword iii

MINIMAX AND GAME PROBLEMS
Chairman: J. B. Cruz, Jr., University of Illinois at Urbana

ON TWO-PERSON, COOPERATIVE GAMES 1
H. B. Poza and R. L. Kashyap, Purdue University

A GLOBAL OPTIMIZATION ALGORITHM 3
R. W. Becker and G. V. Lago, University of Missouri at Columbia

SOLUTION OF THE CONVEX MINIMAX PROBLEM BY THE NEWTON-RAPHSON METHOD 13
J. Medanic, "Mihailo Pupin" Institute

CONDITIONS FOR A MINIMAX OPTIMUM 23
J. W.Bandler, McMaster University

COMMUNICATION SYSTEMS
Chairman: J. B. Thomas, Princeton University

AN ANALYTICAL METHOD FOR FINDING THE INTERVAL PROBABILITY DENSITY OF TWO INTERLEAVED PULSE TRAINS. 31
J. H. O'Hara, Department of Defense

A MODEL FOR PREDICTING PERFORMANCE OF LOW DATA RATE PHASE-LOCKED-LOOP SYSTEMS. 40
T. L. Stewart, Bradley University and K. R. Rao, University of Texas

METHODS OF REDUCING RADAR MULTIPATH TRACKING ERRORS. 48
P. Z. Peebles, Jr., University of Tennessee

SIDELOBE SUPPRESSION IN CHIRP RADAR SYSTEMS. 58
S. Honickman, RCA Laboratories

CODING THEORY
Chairman: F. P. Preparata, University of Illinois at Urbana

MULTIPLE EXTENSION OF THE MINIMUM DISTANCE OF BCH CODES. . . . 68
G. I. Davida, University of Wisconsin-Milwaukee

ON SHORTENED REED-MULLER CODES 71
C. L. Chen, University of Illinois at Urbana and
S. Lin, University of Hawaii

PROBABILISTIC GENERATION, TRANSMISSION AND SYNTACTICAL DECODING OF CONTEXT-FREE PROGRAMMED LANGUAGES. 73
C. R. Souza, University of Notre Dame and
R. A. Scholtz, Universityof Southern California

A CLASS OF MAJORITY DECODABLE MULTIPLE SOLID-BURST-ERROR-CORRECTING CYCLIC CODES. 83
H. T. Hsu, The Pennsylvania State University

ESTIMATION AND IDENTIFICATION OF SYSTEMS
Chairman: N. Wax, University of Illinois at Urbana

ON REDUCING LINEAR DISCRETE SYSTEM STATE ESTIMATION COMPUTER REQUIREMENTS. 90
R. Priemer, University of Illinois, Chicago Circle

KALMAN FILTER FORMULATION OF AN ALTITUDE RATE ESTIMATOR FOR AIRCRAFT FLYING AUTOMATIC TERRAIN FOLLOWING. 92
G. W. Carlock, General Dynamics

THE APPLICATION OF QUADRATIC PROGRAMMING IN ADAPTIVE FILTERING. 101
W. Lochrie, North American Rockwell, Anaheim, and
D. Isaacs, University of California, Irvine

CONSTRAINED OPTIMIZATION OF NONLINEAR FILTERS AND SYSTEMS ON AN APPROPRIATE SYSTEM SPACE 111
N. U. Ahmed, University of Ottawa

ON OPTIMAL NONLINEAR ESTIMATION. 121
J. T. Lo, Stanford University

SYSTEM IDENTIFICATION BY QUASILINEARIZATION USING WEIGHTING COEFFICIENTS 131
C. K. Alexander and G. V. S. Raju, Ohio University

LINEAR SYSTEMS
Chairman: Major J. P. Hyde, Frank J. Seiler Research Laboratory, USAF Academy

EIGENFUNCTION PROPOGATION IN INTERCONNECTED SYSTEMS. 141
L. Kerschberg, PUC, Dept. of Informatica,
Rio de Janeiro, Brasil and W. W. Happ, COPPE-URRJ,
Rio de Janeiro, Brasil

INFLATION IN LINEAR DYNAMICAL SYSTEMS 151
M. K. Sain, University of Notre Dame

A SIMPLIFICATION OF THE STATE REGULATOR PROBLEM. 161
W. L. McDaniel, Jr., Mississippi State University and
V. L. Dugan, Sandia Laboratory

SENSITIVITY MINIMIZATION IN THE DESIGN OF LINEAR MULTIVARIABLE FEEDBACK SYSTEMS 163
M. G. Strintzis, University of Pittsburgh and
B. Liu, Princeton University

TIME OPTIMAL LINEAR TIME-LAG SYSTEMS WITH MULTIPLE CONSTRAINTS ON THE CONTROL FUNCTION. 173
D. H. Chyung, The University of Iowa

COMPUTATION OF GENERALIZED INVERSE AND MATRIX RANK 176
N. N. Gupta, Wright State University

DISTRIBUTED SYSTEMS
Chairman: W. R. Perkins, University of Illinois at Urbana

A DIRECT SEARCH FOR OPTIMAL CONTROL OF DISTRIBUTED SYSTEMS . . 178
A. J. Faulstich, Jr., Naval Ordnance Laboratory and
H. C. Khatri, Catholic University of America

OPTIMIZATION OF PARABOLIC AND HYPERBOLIC DISTRIBUTED SYSTEMS VIA THE CONJUGATE GRADIENT METHOD. 200
D. E. Cornick and A. N. Michel, Iowa State University

ITERATIVE MINIMIZATION IN HILBERT SPACE OF QUADRATIC FUNCTIONALS WITH NONUNIQUE MINIMA. 210
M. Vidyasagra, Sir George Williams University

OPTIMAL CONTROL WITH A BOUNDED ENERGY INPUT DERIVATIVE FOR A CLASS OF DISTRIBUTED SYSTEMS 219
Y. Yavin, Virginia Polytechnic Institute and State University

BOUNDED ENERGY OPTIMAL CONTROL OF TRANSVERSE VIBRATIONS OF A BEAM. 228
Y. Yavin, Virginia Polytechnic Institute and State University and Y. Rasis, Technion, Israel Institute of Technology, Haifa, Israel

A DISTRIBUTED PARAMETER NETWORK APPROACH TO THE ANALYSIS OF ESTUARY POLLUTION. 238
B. P. Butz and R.Fischl, Drexel University

NETWORKS I

Chairman: S. P. Chan, University of Santa Clara

ON THE DESIGN OF A FOUR-PORT DIRECTIONAL FILTER. 248
M. Hibino, Nippon Electric Co., Ltd., Kawasaki, Japan

MINIMAL-CAPACITOR SYNTHESIS OF TRANSFORMERLESS RC n-PORT ON ONE-TREE PORT-STRUCTURES. 258
D. J. Schmidlin and F. J. Lupo, New York University

AN ALGORITHM FOR THE IDENTIFICATION AND AUGMENTATION OF RESISTIVE IRREGULAR NETWORKS 267
J. B. Oso, T. E. Stern, and H. E. Meadows, Columbia University

A NOTE ON A CURIOUS PROPERTY OF THE EQUATIONS OF NONLINEAR NETWORKS CONTAINING TRANSISTORS. 276
D. Mitra, I. W. Sandberg and B. Gopinath, Bell Telephone Laboratories, Inc.

A-INVERSE FORM OF STATE EQUATION AND TRANSFER FUNCTION OF LINEAR NETWORK. 284
Y. Ishizaki, Nippon Electric Co., Ltd., Kawasaki, Japan

SOME NEW TOPOLOGICAL PROPERTIES OF OPEN CIRCUIT IMPEDANCE MATRICES 294
A. Y. Lee, Michigan Technological University and D. P. Brown, University of Wisconsin

SPECIAL SESSION ON SINGULAR PERTURBATIONS

Organizer and Chairman: Petar Kokotović, University of Illinois at Urbana

STRUCTURE OF DECAYING SOLUTIONS OF SINGULAR PERTURBATION PROBLEMS 303
F. Hoppensteadt, New York University

SINGULAR PERTURBATION THEORY OF LINEAR STATE REGULATORS. . . . 310
P. V. Kokotović and R. A. Yackel, University of Illinois at Urbana

SINGULAR PERTURBATION OF A NONLINEAR BOUNDARY VALUE PROBLEM. 322
R. E. O'Malley, Jr., New York University

ON A CLASS OF SINGULARLY PERTURBED TWO POINT BOUNDARY VALUE PROBLEMS 331
C. R. Hadlock, Amherst College

SINGULARLY PERTURBED BOUNDARY VALUE PROBLEMS WITH INTERIOR TURNING POINTS. 340
W. A. Harris, Jr., University of Southern California

STOCHASTIC CONTROL FOR SMALL NOISE INTENSITIES 347
W. H. Fleming, Brown University

SINGULAR PERTURBATIONS AND MINIMUM FUEL SPACE TRAJECTORIES . . 348
T. N. Edelbaum, MIT Charles Stark Draper Laboratory

ANALYSIS OF SINGULAR PERTURBATIONS BY THE METHOD OF GENERALIZED MULTIPLE SCALES. 357
R. V. Ramnath, Electronic Associates, Inc.

SINGULAR PERTURBATION DESIGN OF A CLASS OF TIME-LAG SYSTEMS. . 367
P. Sannuti and P. B. Reddy, Rutgers University

NETWORKS II
Chairman: W. Mayeda, University of Illinois at Urbana

SENSITIVITIES IN TERMS OF WAVE VARIABLES 379
J. W. Bandler, McMaster University and
R. E. Seviora, University of Toronto

SOLVING THE FELDTKELLER EQUATION 388
T. G. Marshall, Jr., Rutgers University

NECESSARY AND SUFFICIENT CONDITIONS FOR THE COMPACTNESS OF AN RC TWO-PORT NETWORK 399
M. J. Massaro, IIT Research Institute and
S. D. Bedrosian, University of Pennsylvania

LC TRANSFER FUNCTION SYNTHESIS BY STATE SPACE TECHNIQUES . . . 408
T. Hua and E. R. Fowler, University of Missouri-Rolla

SOME CHARACTERISTICS OF A COMMUTATED NETWORK FILTER. 410
T. R. Nilsen, Royal Norwegian Navy and G. J. Thaler, Naval Postgraduate School

DISPLAY OF FUNCTIONAL RELATIONSHIPS BETWEEN INTEGER SELECTION CRITERIA 412
W. W. Happ, U. S. Army Corps of Engineers and University of Illinois at Urbana

PATTERN RECOGNITION AND INFORMATION RETRIEVAL
Chairman: W. J. Poppelbaum, University of Illinois at Urbana

A RELATIONAL DATA STRUCTURE FOR DOCUMENT RETRIEVAL 417
Y. H. Chang, R. T. Chien, N. M. Esser, Jr., and F. A. Stahl, University of Illinois at Urbana

OPTIMAL DELAYED DECISION PROCEDURES FOR UNSUPERVISED PATTERN RECOGNITION SYSTEMS. 424
A. G. Jaffer and S. C. Gupta, SMU Institute of Technology

ON THE FIELD CLASSIFICATION TECHNIQUES 427
T. Huang, Purdue University

A PATTERN RECOGNIZING STUDY OF PALM READING. 435
M. Oda, Automatic Control Laboratory, Nagoya University, Nagoya, Japan; B. F. Womack, The University of Texas at Austin; K. Tsubouchi, Mitsui & Co., Ltd., Tokyo, Japan

PATTERN PROCESSING USING MULTILEVEL SYSTEMS. 445
G. H. Granlund, Massachusetts Institute of Technology

SAMPLED-DATA SYSTEMS
Chairman: B. C. Kuo, University of Illinois at Urbana

SAMPLED-DATA CONTROL SYSTEM DESIGN IN THE PARAMETER PLANE. . . 454
S. M. Seltzer, George C. Marshall Space Flight Center, NASA

INITIAL CONDITION ZONES FOR IPFM SYSTEM STABILITY. 464
W. J. Guy, Jr., Lafayette College and M. Kurland and A. Meyer, Newark College of Engineering

Z TRANSFORMS AS HYPERBOLIC TANGENT TRANSFORMATIONS 466
S. A. White, North American Rockwell Corp.

STABILIZATION OF A CLASS OF NONLINEAR SAMPLED-DATA SYSTEMS USING A LINEAR SAMPLED-DATA SYSTEM ESTIMATOR 474
M. I. Younis and W. G. Vogt, University of Pittsburgh

QUANTIZATION ERROR IN CLOSED LOOP DIGITAL SYSTEMS PROGRAMMED IN THE CANONICAL FORM 484
M. H. Fong, J. L. Lowry, J. S.Boland, III,
Auburn University

DIGITAL SYSTEMS AND AUTOMATA THEORY
Chairman: D. R. Schertz, Bradley University,
Peoria, Illinois

ON PERIODICITY OF SEQUENTIAL MACHINES. 486
R. K. Guha and R. T. Yeh, The University of
Texas at Austin

CLOSURE PROPERTY OF LINEAR SEQUENTIAL MACHINES 496
K. C. Wang, Washington State University

A METHOD FOR FINDING MAXIMUM COMPATIBILITY CLASSES FOR AN INCOMPLETELY SPECIFIED FSS. 503
M. A. Tapia, Georgia Institute of Technology

SOME NEW RESULTS CONCERNING SEPARATION THEORY OF GRAPHS. . . 505
D. C. Plisch, Bell Telephone Laboratories and
D. P. Brown, University of Wisconsin

ON THE PLANAR DECOMPOSITION OF INTERCONNECTION NETWORKS FOR COMPUTER-AIDED ELECTRONIC DESIGN 515
G. Bracchi, Istituto di Elettrotecnica ed
Elettronica-Politecnico di Milano, Italy

MEMORY ACCESS CONFLICT RESOLUTION IN MULTIPROCESSING 525
J. M. Daniel and J. D. Irwin, Auburn University

NONLINEAR SYSTEMS
Chairman: R. E. Klein, University of Illinois
at Urbana

OPTIMAL CONTROL FOR A CLASS OF NONLINEAR HEREDITARY DYNAMICAL SYSTEMS. 526
N. U. Ahmed, University of Ottawa

PULSE MODULATED OPTIMAL CONTROLLERS FOR TIME LAG SYSTEMS . . 536
J. Waninski and A. G. Vacroux, Illinois Institute of
Technology

SENSITIVITY REDUCTION IN OPTIMAL CONTROL SYSTEMS 545
N. K. Sinha and S.-H. Dai, McMaster University

LIMIT CYCLES IN INSTRUMENT SERVOS. 552
K. W. Han, Chung-Shan Institute and National
Chiao-Tung University and G. J. Thaler, Naval
Postgraduate School

ON A CLAMPED ATTITUDE-CONTROL MODE FOR AN ORBITING SPACE VEHICLE. 562
D. W. Sutherlin and J. S. Boland, III, Auburn University

A MANIFOLD-IMBEDDING SOLUTION OF OPTIMAL CONTROL PROBLEMS 564
G. D'Ans, Y. Hontoir, and M. Jamshidi, University of Illinois at Urbana

STOCHASTIC SYSTEMS
Chairman: A. H. Haddad, University of Illinois at Urbana

NONLINEAR EXTENSIONS OF A LIMIT THEOREM. 576
R. Lugannani, Princeton University

PROBABILITY DENSITY TECHNIQUES FOR INTEGRAL QUADRATIC FORMS. 587
J. K. Murdock and S. R. Liberty, University of Notre Dame

ON THE APPLICATIONS OF RANDOM SIGNALS WITH MULTIPLICITY ONE. 589
A. Ephredmides and J. B. Thomas, Princeton University

SOME CLASSES OF ANALYTIC RANDOM PROCESSES. 599
K. C. Daly, Air Force Unit Post Office, Department of the Air Force and J. B. Thomas, Princeton University

ON THE REPRESENTATION OF WEAKLY CONTINUOUS STOCHASTIC PROCESSES. 610
S. Cambanis, University of North Carolina and E. Masry, University of California

COMPUTER-AIDED DESIGNS
Chairman: A. H. Haddad, University of Illinois at Urbana

COMPUTER AIDED DESIGN WITH A MINI-COMPUTER 619
T. D. MacCoun, Quindar Electronics, Inc.

THE OPTIMIZATION OF NUMERICAL MODELS OF SEMICONDUCTOR DEVICES. 625
L. C. McAfee, Jr., The University of Michigan

MINIMIZATION OF SENSITIVITY OF ELECTRICAL NETWORKS OVER A FREQUENCY INTERVAL 635
V. K. Manaktala, Bell Telephone Laboratories and
G. L. Kelly, University of Kansas

INTEGRATING TOPOLOGICAL AND NUMERICAL METHODS FOR SEMI-SYMBOLIC NETWORK ANALYSIS 646
G. E. Alderson and P. M. Lin, Purdue University

SYMBOLIC NASAP FOR LINEAR CIRCUIT DESIGN 655
R. S. Schwartz, Northeastern University and
Alford Manufacturing Company

ON TWO-PERSON, COOPERATIVE GAMES *

H. B. Poza and R. L. Kashyap
School of Electrical Engineering
Purdue University
Lafayette, Indiana 47907

ABSTRACT

A method is presented here for the solution of a two-person, dynamic, cooperative game problem for finite time and uncertain measurements. For the case of quadratic criterion, we obtain the optimal control and transmission signal for each of the players. The design of optimal filters for information gathering is also considered.

PROBLEM TO BE SOLVED

The basic idea of the problem being considered is as follows. We wish to investigate the problem of two players, A and B, trying to match their states. Player A sends a message to B concerning his own state. Player B takes a noisy measurement of this message and proceeds to make his own estimate of the state of A. He will then use this estimate to choose a control function which will carry him as close as possible to the estimated state of A. This action taken by B is subject to error (i.e., plant B is subject to plant noise). Player B's action is restricted by a control energy constraint.

Player B then sends the information regarding his new state to A. This information is received inexactly by player A, due to a noisy feedback channel, forcing A to make an estimate of the new state of B. According to this estimate, A chooses the control function which brings him the closest possible to the estimated state of B. Player A's action is also subject to error. As in the case of player B, A's control energy is constrained. Player A will then send B the information regarding his new state and the process is repeated once again. To make the problem meaningful we will also constrain the signal energies of both players.

Our purpose is to obtain the best control functions and transmission signals for both players A and B so as to minimize the final state error between the two players, subject to certain energy constraints. Consider then the system defined by equations

$$x_A(i+1) = Ax_A(i)+Bu_A(i)+\xi_A(i); \quad x_B(i+1) = Cx_B(i)+Du_B(i)+\xi_B(i) \tag{1}$$

where A, B, C and D are known matrices,

$$z_A(i) = v_B(i) + \eta_B(i); \quad z_B(i) = v_A(i) + \eta_A(i) \tag{2}$$

where we assume that $E\{\eta_A(i)\eta_A^T(j)\} = Q_i\delta_{ij}$; $E\{\eta_B(i)\eta_B^T(j)\} = S_i\delta_{ij}$; $E\{\xi_A(i)\xi_A^T(j)\} = R_i\delta_{ij}$; $E\{\xi_B\xi_B^T(i)\} = T_i\delta_{ij}$; and $\eta_A(i)$, $\eta_B(i)$, $\xi_A(i)$ and $\xi_B(i)$ are mutually independent. For the problem in question we will also assume that the signal functions are of linear form; i.e., $v_A(i) = K_A(i)x_A(i)$ and $v_B(i) = K_B(i)x_B(i)$.

We are interested in obtaining the optimal control pair $\{u_A^*(i), u_B^*(i)\}$ and optimal signal pair $\{v_A^*(i), v_B^*(i)\}$, which will minimize $J = E\{\|x_A(N)-x_B(N)\|^2\}$ subject to the energy constraints $E\sum_{i=0}^{N-1}\|v_A(i+1)\|^2 \leq P_o$; $E\sum_{i=0}^{N-1}\|v_B(i+1)\|^2 \leq P_1$; $E\sum_{i=0}^{N-1}\|u_A(i)\|^2 \leq P_2$; $E\sum_{i=0}^{N-1}\|u_B(i)\|^2 \leq P_3$.

---

*This work was partially supported by NASA under Grant NGR 15-005-021, by NSF under Grant GK-1970, and by the AFOSR under Grant 69-1776.

SOLUTION OF THE PROBLEM

The solution of the problem can be divided into three parts:

(a) Optimization of Control Functions

Using the theory of dynamic programming we can obtain the optimal control functions

$$u_A^*(i) = -\Phi_i\{A^{N-i}x_A(i)-C^{N-i}\hat{x}_B(i)\}; \quad u_B^*(i) = -\psi_i\{C^{N-i}x_B(i)-A^{N-i}\hat{x}_A(i)\} \qquad (3)$$

where Φ_i and ψ_i are functions of the plant parameters and $\hat{x}_A(i)$=estimate of A's state by B = $E[x_A(i)|Z_B(i)]$; $\hat{x}_B(i)$=estimate of B's state by A = $E[x_B(i)|Z_A(i)]$, where $Z_A(i)$ and $Z_B(i)$ represent the complete history available to players A and B respectively, up to time i.

It is worthwhile to note here that for the problem in question, it can be readily shown that

$$\left.\begin{aligned} E[E\{x_A(i+1)|Z_B(i)\}|Z_A(i)] &= E[x_A(i+1)|Z_A(i)] \\ E[E\{x_B(i+1)|Z_A(i)\}|Z_B(i)] &= E[x_B(i+1)|Z_B(i)] \end{aligned}\right\} . \qquad (4)$$

Equations (4) are used in obtaining equations (3).

(b) Design of Optimal Filters

Substituting equations (3) into equations (1), and invoking the Kalman-Bucy filtering theory, we obtain the estimate equations to be

$$\left.\begin{aligned} \hat{x}_A(i+1) &= A^*(i)\hat{x}_A(i) + B^*(i)x_B(i) + C_B(i+1)\omega_B(i+1) \\ \hat{x}_B(i+1) &= C^*(i)\hat{x}_B(i) + D^*(i)x_A(i) + C_A(i+1)\omega_A(i+1) \end{aligned}\right\} \qquad (5)$$

where $A^*(i)$, $B^*(i)$, $C^*(i)$ and $D^*(i)$ are functions of the plant parameters, and

$$\left.\begin{aligned} \omega_A(i+1) &= z_A(i+1) - K_B(i+1)C^*(i)\hat{x}_B(i) - K_B(i+1)D^*(i)x_A(i) \\ \omega_B(i+1) &= z_B(i+1) - K_A(i+1)A^*(i)\hat{x}_A(i) - K_A(i+1)B^*(i)x_B(i) \end{aligned}\right\} . \qquad (6)$$

$C_A(i)$ and $C_B(i)$ are the filter gain constants. It can also be shown in the same fashion as equations (4) that for this problem

$$E[E\{x_A(i)|Z_B(i)\}|Z_A(i)] = x_A(i); \quad E[E\{x_B(i)|Z_A(i)\}|Z_B(i)] = x_B(i). \qquad (7)$$

Equations (7) are used in obtaining equations (5).

(c) Optimization of Signal Functions

Since the signal functions have been assumed to be linear, the problem of obtaining the optimal signal functions clearly reduces to the problem of finding the optimal values of the gains $K_A(i)$ and $K_B(i)$, so as to minimize J. This problem can easily be solved by using any of the standard techniques of optimal control theory.

REFERENCES

1. Kalman, R. E., and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," Journal Basic Engr., Vol. 83, March 1961.
2. Bryson, A. E., Jr., and Y. C. Ho, Optimization, Estimation, and Control, Blaisdell Publ. Co., New York, N.Y., 1968.
3. Kashyap, R. L., "Stochastic Cooperative Games with Applications," unpublished manuscript.
4. Tung, F., "Linear Control Theory Applied to Interplanetary Guidance," IEEE Trans. on Automatic Control, Vol. AC-9, pp. 82-89.

A GLOBAL OPTIMIZATION ALGORITHM

Ronald W. Becker, G. V. Lago
University of Missouri at Columbia
Columbia, Missouri

ABSTRACT

A global optimization algorithm using a Mode -Seeking Program and Monte Carlo analysis is described. The parameters of a function to be minimized can be varied at random over any desired range, subject to any restrictions. The algorithm is especially adaptable to circuit theory design.

I. INTRODUCTION

The advent of the digital computer has permitted the solutions of many optimization problems whose solutions would have been impractical otherwise. Gradient methods, such as Fletcher and Powell's method [1], often require many derivative calculations and function evaluations in each step of the iterative process. Alternatives to these gradient methods are methods which require only function evaluations. Rosenbrock [2], Hooke and Jeeves [3], and Nelder and Mead [4] (Simplex Method) have presented methods which require only function evaluations. The method presented in this paper is an extension of random search methods as discussed by Brooks [5]. The latter methods are sometimes referred to as direct search methods, whereas the gradient methods are also referred to as descent methods. Descent methods can require many derivative and function evaluations, while direct methods require many more function evaluations but no derivative calculations.

The random search procedures to be discussed will be concerned with minimization of a function E, which, in the case of network design, includes some design criteria. Usually E will be a measure of the error between a desired response F and a randomly generated response F'. E may be a least squares error, a Chebyshev error or a coefficient matching error function. F' will generally be a function of two sets of variables: first, the set of parameters, X_i, i=1,---,k, to be randomly generated in order to minimize E; and, secondly, the set of points, Y_j,j=1,---,w, where F and F' are measured. For network design the set Y_j may correspond to a set of frequencies at which F and F' are measured.

In any random search method trials are made at combinations of parameters selected in some random fashion. The estimate of the optimum parameter combination is simply the combination of parameters which produced the lowest value for the error function. This process can also be classified as a Monte Carlo method [6].

Brooks has reviewed several random search techniques: the simple random method, the stratified random method and the creeping random method [7]. The simple random method basically makes trials at points X_i selected at random throughout the whole experimental region. The trial with the lowest error value is called the optimum solution. In the stratified random sampling method the experimental region is divided up into subregions by a regular grid. All of these rectangular subregions are of equal size and the number of subregions is made to correspond to the number of desired trials. A trial is made at a randomly selected point within each subregion. The optimum solution is that combination of parameters which has the lowest error evaluation. The creeping random method is different from the previous two methods in that it is a sequential random procedure. A guess is made initially and then a set of random trials is made around the initial guess. The random trials may be selected from a normal distribution that is centered on the initial guess. The optimal point of the set of trials is retained as the center of another hypothetical normal distribution, from which another set of random trials is sampled. The procedure is repeated, and the variance of the normal distribution is diminished each time, thereby converging to a local minimum.

II. RANDOM SEARCH OPTIMIZATION BY MODE-SEEKING

A new algorithm has been developed to improve upon simple random methods by implementing a mode-seeking program. The algorithm is a two or more step procedure depending on the desired degree of minimization.

Step one is essentially a simple random search over the entire experimental parameter region. Instead of retaining only the point with the minimum error as the unaltered simple random method does, the proposed method retains a predetermined number of points which have the lowest error values of all the original trials. If the error function were unimodal over the entire experimental parameter region, then the set of retained points would be centered about the absolute minimum of the unimodal function. The points could be said to form a cluster about the minimum. The compactness of the cluster would be dependent upon the number of original random trials and the number of points retained. If the error function were bimodal and the modes were symmetric with equal absolute minimums, then the retained points would geographically form two clusters. Suppose the error function were bimodal, but one mode had an absolute minimum much lower than the other mode. It is still likely that the retained points would form two clusters, though most of the points would probably be located around the lowest absolute valued mode, particularly if the two modes had similar contours. These ideas can be extended to include functions with many modes, functions with modes within modes, and functions which have spike-shaped modes. Irregular error functions will naturally require more initial trials to obtain a good set of retained points and, therefore, clusters which are representative of modes.

The retained points are grouped into clusters by an algorithm based on the geographical location of the points with respect to each other. Details of the particular algorithm will be discussed later. Since this program is designed to find clusters which correspond to modes, the program will be called a Mode-Seeking Program.

At this point the clusters found by the Mode-Seeking Program are graded and rated. The clusters are graded by searching in each cluster for the retained point which has the lowest error value. The clusters are then rated according to the points found in the grading process. The best cluster would be the cluster containing the point with the lowest error value. Consider again the bimodal error function with unequal absolute modal minimums. Assume that a proper number of initial trials are made and that the retained points are classified into two clusters. Then, with a very high probability, the cluster located around the lower valued mode would be rated as the better cluster.

A subregion of the original experimental region is constructed to contain all of the retained points of each cluster. The bimodal error function would have two subregions, one for each cluster. In some instances it might be advantageous to enlarge the subregions somewhat to insure that the absolute minimums are contained within the subregions. The random search method has essentially been extended to create and store information concerning the topology of the error function. Additional steps will use this topological information to locate the minimum of the error function. This concludes the first step of the proposed random search method using a mode-seeking program.

Step two of the proposed random search method treats each subregion found in the first step as an experimental region upon which a new random search is undertaken. For example, assume that two clusters were found for the bimodal error function by the first step, then the second step takes a set of random trials on each subregion formed by the two clusters. The second step proceeds as in the first step: it retains a certain number of the points with the lowest error values, clusters the points, rates the clusters, and finally creates sub-subregions. It is important to specify that the second step performs a random search first on the best cluster, next on the second best cluster and so on. Theoretically, the absolute minimum of the error function should lie somewhere within the best cluster. Consequently, it would appear that the second step should only perform a random search on the best cluster. In practice, however, it is a good idea to include the second best cluster and possibly the third to allow for probability discrepancies. Much computer time can be saved by limiting the number of subregions searched by the second step.

The pattern of the proposed random search method with a mode-seeking program should be evident. If necessary, the third step and possibly higher steps may be taken in a similar manner to the second step. Each additional step should reduce the subregions to smaller and smaller intervals and in some cases, to the exact absolute minimum. A terminating error bound based on the desired level of solution accuracy may also be provided as the point below which no further improvement is necessary.

Global Optimization and Constraints

The proposed random search method with a mode-seeking program is a global optimization algorithm. The initial experimental region under investigation may be as large as desired and have as many parameters as desired. Recall that the parameters of F were $X_i, i=1,---,k$, where k is the order of the parameter space. The parameters may also have any constraints placed on them since the parameters can be generated very easily to satisfy the constraints.

Such an optimization procedure has many advantages. Gradient methods usually work well and fast; however, these methods necessarily require that the error function be unimodal over a region. The proposed random search method would work especially well under similar circumstances. If the error function is multi-modal, such as in typical network problems, and gradient optimization methods are applied, the result will probably be a local minimum of disputable merit. Successful gradient techniques rely heavily on the initial point from which the optimization proceeds. A multi-modal error function would be solvable by the proposed random search technique. Such a problem may take many trials and several algorithm steps in the process. The simple random search method and the stratified random search method might eventually find the global minimum, but they would certainly take many more trials than the proposed random search algorithm. The creeping random method, as stated before, would only find a local minimum for a multi-modal error function.

Parameter scaling is often necessary with gradient methods. In particular, the element values of a network must often be normalized to use Fletcher and Powell's method [8]. Parameter scaling, on the other hand, is not necessary with the proposed random search algorithm.

The proposed random search method will also work on problems with nonlinearities and discontinuities. These, of course, must be incorporated within the response functions F and F' and the error function E. If this can be done, then the proposed random search algorithm should find the optimum solution.

The circuit designer may also incorporate another novel type of constraint on the parameter values. Suppose the designer wishes to use only commercially available resistors, capacitors, inductors, transistors, and other components. It would be possible to preload the experimental region with commercial values as parameter points. The end result would be an optimum set of available components. Gradient optimization techniques, on the other hand, usually result in a set of extremely precise parameter values with little likelihood of locating or making the exact element value as specified.

Number of Trials Required

A sufficient number of trials must be taken if the proposed random search method can be expected to be meaningful. The researcher using this method must estimate before running the program the number of trials needed for the initial step and the number of trials for succeeding steps.

Recall the concept of the subregion or subregions as related to the entire experimental region. The first step of the proposed random search algorithm defines a subregion (subregions), in the form of clusters, where the optimum solution exists. Some of the initial random trial points, therefore, must fall within the subregion (subregions) where they will be retained. Let P denote the relative size of the subregion or sum of subregions with respect to the entire experimental region. It may be seen that P is also the probability that a randomly selected trial point will fall within the desired subregion or one of the subregions. The distribution of random trials is simply a binomial distribution.

Let m denote a number of retained points and n the number of random trials. Then from the binomial distribution, the probability, S', of m successes in n trials is

$$S' = \frac{n!}{m!(n-m)!}P^m(1-P)^{n-m}$$

It is possible to specify S', P, and m and then calculate n. However, due to the factorials, finding n can be very difficult. Since any extra trials may only reduce the size of the subregion (subregions) more than expected, an easier estimate of the number of trials was developed. Let S denote the probability of at least one trial falling in the subregion out of n trials. Then

$$S = 1 - (1-P)^n ,$$

which is one minus the probability of no points falling in the subregion. Solving for n,

$$n = \log(1-S)/\log(1-P) .$$

If n trials have probability S of at least one successful trial, then the product mn trials will have a probability greater than S of at least m successful trials. The product mn is an over-estimate but any excessive trials are also useful.

The proposed random search algorithm has been programmed so that the researcher must provide the error function E, the parameter generation scheme, the number of trials for each step, and the number of retained points for each step before executing the program. The method could be improved, however, by programming the algorithm into a conversational mode on any capable computer. A user could monitor the progress of the method at suitable points and, if necessary, the user could direct more trials to be taken and perhaps fewer or more points to be retained. One could also eliminate searching through doubtful subregions and perhaps intensify the search in other subregions. A researcher experienced in optimization would certainly appreciate the conversational mode version of the proposed algorithm.

Random Parameter Generation

Any number of random parameter generating schemes can be implemented within the proposed algorithm. There is no one best scheme suited for all situations. Experience has shown that special-purpose generating schemes can save considerable computer time. The only limit on the number and type of generating schemes is the ingenuity of the program user.

Basically all generating schemes on a digital computer use a multiplicative congruential method based on the relation (involving non-negative integers k) $Z_{i+1} = aZ_i \pmod{k}$ $(0 \le Z < k)$, which means that the expression aZ_i is to be divided by k [9]. The constant a is chosen to provide speed, a long period, and good statistical behavior. Since only the remainder is retained on division by k, the period cannot be greater than k. Therefore, k is usually chosen to be one more than the largest computer integer. It is then unnecessary to divide by k, since the computer cannot register any product larger than k-1. This automatically provides the remainder. The set of Z's are the random integers which are then converted by additional user-supplied schemes into various sets of random parameters. There are numerous tests available, such as the frequency, serial, gap and poker tests, which test the randomness of a set of numbers [10]. For the purposes of the proposed random search method, a uniform distribution of the random parameters is usually best.

As an illustration, suppose random integers are desired between 0 and 10. Suppose also that a program exists which generated the Z's. Then a Fortran IV statement function might be defined as

```
INTGR(Z,J) = Z(mod J).
```

The integers, $0<I<10$, could be produced by

```
I = INTGR(Z,9) + 1.
```

Say for example, the following set of integers need to be produced randomly: 1,2,3,---,10,20,30,---,100,200,300,---,900. A solution would be

```
I = (INTGR(Z,9) + 1)*(10.0**INTGR(Z,3)).
```

Finally, suppose the following set of integers need to be produced randomly: -900,-800,---,-100,-90,-80,---,-10,-9,-8,----1,1,2,---9,10,20,---,90,100, 200,---,900. A solution would be:

```
I = (-1)**INTGR(Z,99)*(INTGR(Z,9) + 1)*(10.0**INTGR(Z,3)).
```

Mode-Seeking Program

The proposed random search method will work well only if the mode-seeking program works well. The mode-seeking program may also be referred to as a clustering program. Basically, the program must input a set of points in K space and output the points in clusters or groups. The points are grouped into clusters according to their geographical location with respect to one another. Unfortunately, the mode-seeking program has no knowledge concerning the number of clusters to expect, so it must be programmed to decide internally how many clusters best fit all of the input points.

The proposed random search method uses a mode-seeking program developed by Bryan [11]. This program performs an estimate of the true positive density function by using potential functions, followed by a hill climbing scheme which finds the modes. The program uses a Mahalonabis distance transformation to allow very large magnitudes between parameters. This may occur for example, when one parameter may be in the micro-farad range and another may be in the kilo-ohm range. Tests have shown that Bryan's mode-seeking program works very well; however, it is too detailed to present in this paper. There are probably additional clustering algorithms available which would work satisfactorily in the proposed random search method.

III. EXAMPLES

Example #1--Four Minimums

A simple but very illustrative example was optimized by the proposed random search method. The example was devised to have four minimums, the solutions of

$$E = \left[|x_1| - 5\right]^2 + \left[|x_2| - 5\right]^2 .$$

It was assumed that the minimums were in the intervals $-10^7<x_1<+10^7$ and $-10^7<x_2<+10^7$, large enough to illustrate the global capacity of the proposed algorithm. A random number generation scheme, similar to the last example in Section II, was written.

The following notation has been developed to simplify example descriptions:

R = set of all points in the initial experimental region,

t_i = set of mn random trials for step i,

r_i = set of m retained points for step i,

and R_a = set of all points in subregion determined by cluster 'a'.

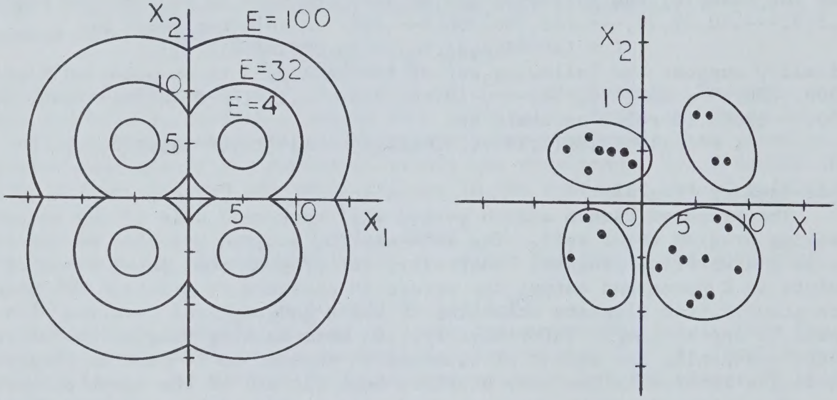
For this example, $R = 126^2$ or 15,876 points since x_1 and x_2 are each defined for X_{ji}; $i = 1,---,126$, $j = 1,2$.

Figure 1.a is a plot of some theoretical E (Error) function contours. The solid lines are constant values of E as indicated.

Suppose a subregion reduction factor of 25 is sought with 95% probability of at least one point falling in this subregion. Then

$$n = \log(1-.95)/\log(1-.04)$$
$$n = 73.$$

Suppose also that 25 points are to be retained. Then mn = 1835, and for purposes of rounding, t_1 is set at 2000 trials. Two thousand trials are then generated at random, E is evaluated at each trial, and the 25 points in r_1 are clustered by the mode-seeking program into four clusters as shown in Figure 1.b. Some of the clusters are centered around the minimums; (-5, -5), (-5,5), (5,-5), and (5,5), while some are not. However, each cluster does contain the proper minimum within its own subregion.



(a) Error Contours - Ex. #1

(b) Clusters - Ex. #1

Figure 1

Step two of the proposed algorithm was run on each of the four subregions with $t_2 = 100$ points and $r_2 = 5$ points. In every subregion examined by step two, the exact minimum, E = 0, was found. Step two was the final step for this example.

This example points out several important aspects of the proposed optimization method: its global capability, the calculation of mn, the power and accuracy of the mode-seeking program, and the relatively small number of trial points needed to find the four absolute minimums.

Example #2--Parameter Constraints

This example was suggested first by Beale [12] and is recognized as a standard test problem for optimization methods. The nonlinear function to be minimized is

$$E = 9-8x_1-6x_2-4x_3+2x_1^2+2x_2^2+x_3^2+2x_1x_2+2x_1x_3$$

subject to the constraints

$$x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,$$

and

$$x_1 + x_2 + 2x_3 \leq 3.$$

The last constraint is active. Beale states that the minimum is E = 1/9 at $x_1 = 4/3$, $x_2 = 7/9$, and $x_3 = 4/9$.

From the statement of the problem, the ranges on the x's are:

$$0 < x_1 < 3$$
$$0 < x_2 < 3$$
$$0 < x_3 < 1.5$$

These values define the initial experimental region. The x's can only be randomly generated to a definite number of decimal places and, in this case, two decimal digits were chosen. The fortran statements to generate the x's are:

```
10 x1 = INTGR(Z,299)*1.0E-2 + .01
   x2 = INTGR(Z,299)*1.0E-2 + .01
   x3 = INTGR(Z,149)*1.0E-2 + .01
   IF ((x1 + x2 + x3).GE. 3.0) GO TO 10
```

where the Z's are randomly generated integers and "INTGR" is a modular arithmetic statement function as previously defined. The active constraint is easily handled by the conditional "GO TO". Obviously, the exact solution cannot be found since 4/3, 7/9, and 4/9 cannot be explicitly expressed in two-place decimal numbers. For this problem, S and P were selected to be .85 and .01 respectively. This should reduce the experimental region by a factor of 100 with 85% probability. If $r_1 = 10$, then t_1 is approximately 2000 trials (R = 13,320,749).

Step one produced one cluster referred to as A, the best point having E = 0.122. The corresponding subregion defined by the cluster is

$$1.15 \le x_1 \le 1.52$$
$$0.62 \le x_2 \le 0.87$$
$$0.32 \le x_3 \le 0.52$$

which has $R_A = 23{,}166$ points. So the first step actually reduced the initial experimental region by a factor of 500 or more.

Step two was run with $t_2 = 1000$ trial points and $r_2 = 5$ retained points. This should also have a 100 to 1 reduction factor. E was reduced to 0.112 and one cluster, AA, was formed. The reduced subregion is

$$1.31 \le x_1 \le 1.37$$
$$0.76 \le x_2 \le 0.79$$
$$0.42 \le x_3 \le 0.46$$

which has $R_{AA} = 240$ points, a reduction of almost exactly 100.

The third step found the best solution possible with two decimal place numbers,

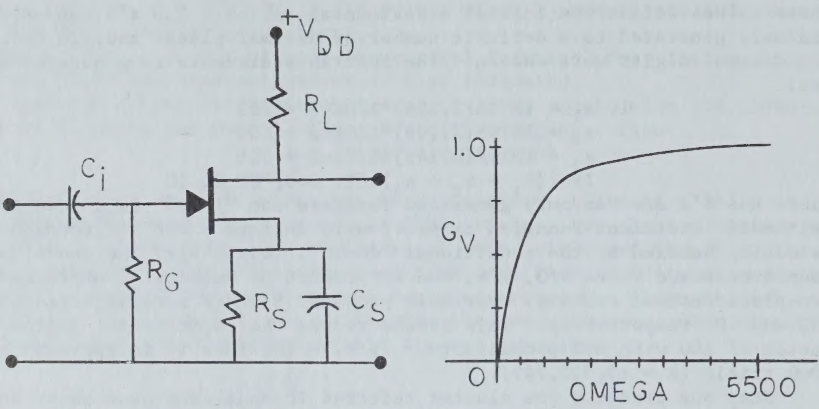
$$x_1 = 1.34$$
$$x_2 = 0.78$$
$$x_3 = 0.44 \quad .$$

E was reduced to .111199 as compared to Beale's 1/9 = .111111---. Example #2 illustrates the method of applying constraints on the parameters and the rapid convergence to an absolute minimum in 3,500 trials. The proposed random search method with a mode-seeking program found the minimum using only .025% of the total number of possible points in R.

Example #3--Field Effect Transistor

This last example was chosen to illustrate the usefulness of the proposed optimization method to a network designer. Consider matching the voltage gain of a Field Effect Transistor amplifier with a desired set of voltage gain points measured at a set of frequencies. Such a problem could be handled with a least squares minimization error function. Figure 2.a shows the circuit of the Field Effect Transistor (FET) and Figure 2.5 shows the desired voltage gain and the twelve sampling radian frequencies. The four parameters to be optimized are: C_i, R_g, R_s, and C_s. Two elements g_{fs} and R_1 were assumed to be constant with values of .001 mho and .1Meg ohm, respectively.

The proposed random search method has a network analysis program incorporated as a sub-program. NASAP was chosen as the network analysis program since it solves for a network's transfer function symbolically. The transfer function may be one of many, for instance a voltage gain, driving point admittance or a transfer impedance. Once NASAP finds the transfer function it is available for further use and repeated calculations without destroying the function.



(a) FET Amplifier (b) FET Voltage Gain

Figure 2

Referring to the previously defined notation, F denotes the set of twelve desired voltage gains at the sampled radian frequencies. F' is the voltage gain function from NASAP calculated at $s=jw_i$, $i=1,---,12$. E is then the least squares error

$$E = \sum_{i=1}^{12} (F_i - F'(X_j, W_i))^2 \qquad j=1,2,3,4.$$

where

$$F' = \frac{(R_g R_s R_l g_{fs} C_i C_s)s^2 + (R_g R_l g_{fs} C_i)s}{(R_g R_s C_i C_s)s^2 + (R_g C_i + R_s C_s + R_g R_s g_{fs} C_i)s + R_s g_{fs} + 1}$$

The random parameter generating statements were written to produce random integers, 9 to a decade, for the following ranges:

$$10^{-9} \le C_i < 10^{-6}$$
$$10^{6} \le R_g < 10^{9}$$
$$10^{3} \le R_s < 10^{6}$$
$$10^{-6} \le C_s < 10^{-3}$$

Each dimension in four-space has 30 points and as a result R = 810,000.

The sets of retained points r_1, r_2, and r_3 were set at 25, 15, and 15 points respectively. The number of trials t_1 was set at 2500 for the first step which should have a reduction factor of at least 30 or 40. The first step resulted in three clusters, denoted by A, B, and C. Most of the points in r_1 fell in the best cluster, A. Each of the 25 retained points in r_1 had C_s equal to $1. \times 10^{-6}$, so in all further steps C_s was considered to be optimized at that value. The sub-regions $R_A + R_B + R_C$ totaled 7,000 points, a reduction factor of over 100.

Step two examined $t_2 = 500$ trials for each of the two best clusters A and B produced by step one. The best cluster of step one produced very low error points while the second best cluster has errors five orders of magnitude higher than the best cluster. Therefore, only the A cluster was considered significant. The second step produced two clusters, AA and AB, from the A cluster of step one. Furthermore, R_s was equal to $1.\text{x}10^4$ for all of the points in r_2; so once again, this is considered to be the optimum value.

The third step was preprogrammed to examine only the best cluster produced by the second step, in this case AA. This step found two clusters, AAA and AAB, both of which had very low error values. The AAA cluster had $E=.18\text{x}10^{-7}$, $C_i=1.\text{x}10^{-8}$, and $R_g=1.\text{x}10^{7}$. The AAB cluster had $E=.21\text{x}10^{-7}$, $C_i=5.\text{x}10^{-9}$, and $R_g=2.\text{x}10^{7}$. Both sets of points are certainly valid optimum solutions and the resulting plot of voltage gain matched the desired curve in Figure 2.b.

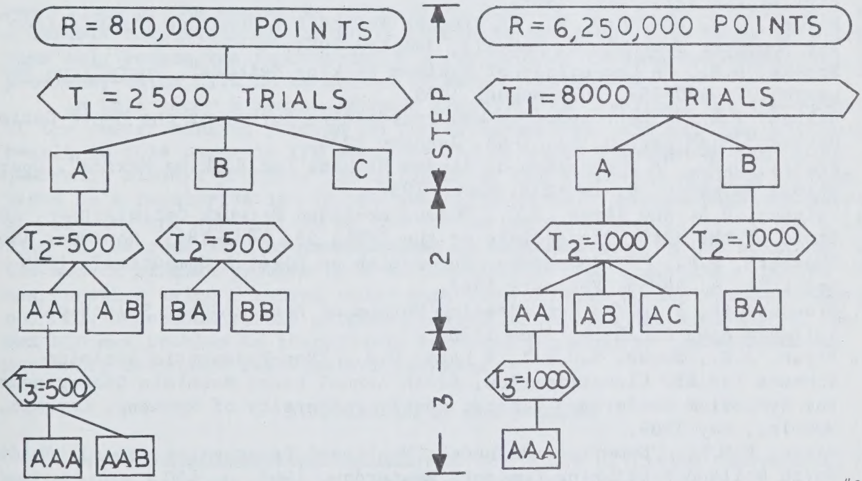
This problem is typical of multi-modal problems in general. The error surface is very irregular and modes often exist within larger modes. Essentially, one mode may appear from a large perspective; however, several modes may appear within the larger mode from a smaller, restricted viewpoint. Figure 3.a summarizes this example with a search chart showing the steps, clusters, and results. Note that the minimums were found using less than .4% of the total number of possible points in R.

This same problem was rerun using a larger initial experimental region to search for minimums other than the two previously found. The expanded ranges on the parameters are:

$$\begin{aligned} 10^{-11} &\le C_i < 10^{-6} \\ 10^{6} &\le R_g < 10^{11} \\ 10^{3} &\le R_s < 10^{8} \\ 10^{-8} &\le C_s < 10^{-3} \end{aligned}$$

which have a total of R=6,250,000 possible points. Figure 3.b shows the details of steps one, two and three. Both the original run and the rerun are very similar up to the output of the third step.

The third step produced one cluster which had four different minimum points: $(R_g, C_i) = (.5\times10^{7}, .2\times10^{-7})$, $(.5\times10^{8}, .2\times10^{-8})$, $(.1\times10^{8}, .1\times10^{-7})$, and $(.2\times10^{9}, .5\times10^{-9})$. Each minimum has a global minimum in its own right; however, upon closer examination the implication is clear. The $R_g C_i$ product is .1 for all minimums and therefore the global minimum occurs whenever R_g and C_i satisfy this condition. The optimization program provided more information about the circuit than expected.



(a) FET Search Chart - Run #1 (b) FET Search Chart - Run #2

Figure 3

IV. SUMMARY

The three previous examples have shown some of the advantages of the proposed random search optimization method using a mode-seeking program. First and foremost is its ability to locate global minimums. A global minimum is never guaranteed with any random method; however, the proposed random method increases the probability of finding the global minimum. In any case the minimum found by this method is usually better than local minimums typically found by gradient methods. Secondly, the proposed method can handle parameter constraints very easily as Example #2 illustrated. In network problems, the optimum solution can be forced to be commercial on-the-shelf components. Thirdly, only error functions need to be calculated, as opposed to error and derivative functions for gradient methods. The proposed method requires many function evaluations, to be sure, but the amount of computer time and success of a solution is the real measure of a program. Fourthly, the proposed method can be programmed to handle non-linear networks, and networks in the time domain, although the examples chosen here do not illustrate this phase. Finally, the proposed method could find a small interval containing the minimum, then a gradient method could find the exact minimum.

The proposed optimization method, as with any optimization method, works better with practical experience. Someone who has a feel for the problem under attack can estimate the number of clusters to expect, number of random trials, and number of steps and, consequently, efficiently utilize the method. The optimization method would ideally be suited for a computer with a conversational mode.

REFERENCES

1. Fletcher, R. and Powell, M.J.D., "A Rapidly Convergent Descent Method for Minimization", The Computer Journal, 6, 163-168, July 1963.
2. Rosenbrock, H.H., "An Automatic Method for Finding the Greatest or Least Value of a Function", The Computer Journal, 3, 175-184, October 1960.
3. Hooke, R. and Jeeves, T.A., "Direct Search Solution of Numerical and Statistical Problems", Journal of the Association for Computing Machinery, 8, 212-229, April 1969.
4. Nelder, J.A. and Mead, R., "A Simplex Method for Function Minimization", The Computer Journal, 7, 308-313, January 1965.
5. Brooks, S.H., "A Comparison of Maximum Seeking Methods", Operations Research, 7, 430-457, July-August 1959.
6. Halton, J.H., "A Retrospective and Prospective Survey of the Monte Carlo Method," SIAM Review, 12, 1-63, January 1970.
7. Brooks, S.H., "A Discussion of Random Methods for Seeking Maxima", Operations Research, 6, 244-251, March 1958.
8. Calahan, D.A. and Temes, G.C., "Computer-Aided Network Optimization, the State of the Art", Proceedings of the IEEE, 55, 1832-1863, November 1967.
9. Chambers, R.P., "Random-Number Generation on Digital Computers", IEEE Spectrum, 4, 48-56, February 1967.
10. Gruenberger, F., "The Test-Passing Method of Random Digit Selection", Software Age, 4, 15-16, June 1970.
11. Bryan, J.K., Dwyer, S.J.III, & Lago, G.V., "Non-Parametric Decision Schemes for EEG Classification, Sixth Annual Rocky Mountain BioEngineering Symposium Conference Record, 56-61, University of Wyoming, Laramie, Wyoming, May 1969.
12. Beale, E.M.L., "Numerical Methods", Nonlinear Programming, ed. J. Abadie North Holland Publishing Company, Amsterdam, 1967, p. 150.

This work is part of a Ph.D. Thesis to be submitted by R. W. Becker.

SOLUTION OF THE CONVEX MINIMAX PROBLEM BY THE NEWTON-RAPHSON METHOD

JURAJ MEDANIC
"Mihailo Pupin" Institute
Belgrade, Yugoslavia

ABSTRACT

Newton-Raphson type algorithm is proposed for the solution of the convex minimax problem. The solution is based on a regularization procedure by which a modified cost function is defined, such that the modification does not change the minimax solution. The modified cost function is such that by the saddle point theorem it is possible to interchange the order of the min and max operations which allows a simple solution of an associated quadratic minimax problem and thereby the construction of a Newton-Raphson algorithm for the convex minimax problem.

INTRODUCTION

Computational aspects of the minimax problem in finite dimensional spaces have recently been considered by a number of authors[1-10]. Generally speaking, three types of algorithms have been proposed for the solution of minimax problems: (i) gradient algorithms, (ii) exchange algorithms, and (iii) elimination algorithms. Elimination algorithms [9] have been found relatively efficient [10] for the class of convex minimax problems to which they are restricted. Gradient [1-4] and exchange [5-8] algorithms are not restricted by the convexity assumption. However, to assure convergence considerable computational burden is generally required; both types of algorithms only reduce the fundamental minimax problem to somewhat simpler minimax problems, which have to be solved in each iteration [4,5].

In this paper a Newton-Raphson algorithm is developed for the solution of the convex minimax problem in finite dimensional spaces. The principal result in this paper is the complete and original solution of the so-called quadratic minimax problem. The solution of the quadratic minimax problem is based on a regularization of the quadratic minimax problem with the purpose of allowing the application of the saddle point theorem [12,13]. The solution of the quadratic minimax problem is then employed in the solution of the convex minimax problem by a Newton-Raphson type algorithm, applied to the max function using a second order expansion related to the result of Demjanov [2]. Numerically, the proposed algorithm reduces the solution of the convex minimax problem to iteratively repeated solution of two related but numerically decoupled maximization problems.

FORMULATION OF THE PROBLEM

Consider the numerical solution of the class of finite dimensional convex minimax problems where the solution is defined by

$$J(x^*,y^*) = \min_{x \in X} \max_{y \in Y} J(x,y) \qquad (1)$$

and where the cost function $J(x,y)$ and the admissible domains X and Y satisfy the following assumptions

(a) For each $y \in Y$, the cost function $J(x,y)$ is strictly convex in X;
(b) The domain X coincides with the n-dimensional euclidean space E_n and Y is a compact subset of the m-dimensional euclidean space E_m;

(c) For each $y \in Y$, the cost function $J(x,y)$ is continuous and possesses continuous forst and second derivatives with respect to x.

For the following development it is convenient to introduce the definitions of the max function $F(x)$, the min function $G(y)$ and the answering sets $Y(x)$ and $X(y)$.

Definition 1. The max function $F(x)$ associated with the cost function $J(x,y)$ is defined by

$$F(x) = \max_{y \in Y} J(x,y). \tag{2}$$

Analogously, the min function $G(y)$ associated with the cost function $J(x,y)$ is defined by

$$G(y) = \min_{x \in X} J(x,y). \tag{3}$$

Definition 2. Given $\underline{x} \in X$, the set of all elements $y \in Y$ maximizing $J(\underline{x},y)$ is denoted by $Y(\underline{x})$ and called the answering set of $\underline{x}$. Analogously, given $\underline{y} \in Y$, the set of all elements $x \in X$ minimizing $J(x,\underline{y})$ is denoted by $X(\underline{y})$ and called the answering set of $\underline{y}$.

It is further assumed that the following assumption is satisfied:

(d) For each bounded $x \in X$, the answering set $Y(x)$ is a finite set.

The necessary and sufficient conditions for x^* to be a solution of the convex minimax problem are well known [1,2,14]: Let

$$g_i(x) = \frac{\partial J}{\partial x}(x,y_i)^T,$$

let T denote transpose, and let γ be a unit direction vector.

Theorem 1. A necessary and sufficient condition for x^* to be the solution of the convex minimax problem is that

$$\min_{\gamma} \max_{y \in Y(x^*)} g_i(x^*)^T \gamma \geq 0, \quad \|\gamma\| = 1. \tag{4}$$

If, in the considered class of problems, the minimax solution is finite, it necessarily belongs to the interior of the domain of x; it has been shown that in such a case the following necessary condition for the minimax solution must be satisfied [11]:

Theorem 2. Let

$$M(x) = \begin{Vmatrix} g_1(x)^T \\ g_2(x)^T \\ \cdot \\ \cdot \\ g_r(x)^T \end{Vmatrix}, \quad y_i \in Y(x), \quad i=1,\ldots,r. \tag{5}$$

A necessary condition for x^* to be the minimax solution is

$$\operatorname{rank} M(x^*) < \min(r,n+1). \tag{6}$$

The above condition implies that at the minimax solution point the vectors $g_i(x^*)$, $i=1,\dots,r$ are linearly dependent. Combining necessary conditions in Theorems 1 and 2, it follows that there exist coefficients c_i^*, $i=1,\dots,r$ such that

$$\sum_{i=1}^{r} c_i^* g_i(x^*) = 0, \tag{7}$$

with c_i^* satisfying

$$\sum_{i=1}^{r} c_i^* = 1, \quad c_i^* \geq 0, \ i=1,\dots,r. \tag{8}$$

AN EXPANSION RESULT AND NEWTON-RAPHSON MINIMIZATION

It has been shown by Demjanov [2] that when the cost function belongs to class C^s with respect to x for each $y \in Y$ it is possible to write an s-th order expansion for the max function $F(x)$ in the vicinity of a nominal point x_k. This series expansion, truncated after the first order term allowed the development of necessary condition (4) and the definition of the negative gradient direction employed in gradient type algorithms.

Truncating after the second order term and carry-ing -out only partially the steps described in [2] for the determination of the final expansion formula, it is possible to write the second order approximation of the max function in the following form:

$$F(x) = F(x_k) + \max_{y_i \in Y(x_k)} \left[g_i(x_k)^T \delta x + \tfrac{1}{2} \delta x^T H_i(x_k) \delta x \right], \tag{9}$$

where $Y(x_k)$ is the answering ser of x_k, $x = x_k + \delta x$, and by definition

$$H_i(x_k) = \frac{\partial^2 J}{\partial x^2}(x_k, y_i). \tag{10}$$

Note that in nonlinear programming problems where $F(x)$ is a sufficiently smooth function the answering set $Y(x_k)$ may be considered a singleton containing the same element for each x_k and the max operation in (9) is trivially performed by dropping the index "i". The philospphy of the Newton-Raphson algorithm in determining the minimum of a sufficiently smooth function is to obtain the optimal step size δx^* by minimizing the right hand side of (9) with respect to δx, with the explicit result

$$\delta x^* = - H(x_k)^{-1} g(x_k). \tag{11}$$

In the minimax problem, generalizing the Newton-Raphson philosophy, one concludes that the optimal δx^* should again be obtained by minimizing the right hand side with respect to δx, which is now equivalent to the solution of the following minimax problem:

$$d^* = \min_{\delta x} \max_{y_i \in Y(x_k)} \left[g_i(x_k)^T \delta x + \tfrac{1}{2} \delta x^T H_i(x_k) \delta x \right]. \tag{12}$$

The efficiency of the Newton-Raphson algorithm in the minimax case evidently depends on the computational burden necessary to solve this associated minimax problem. It is one of the aims of this paper to show that it is possible to transform this minimax problem into a more manageable computational problem.

THE QUADRATIC MINIMAX PROBLEM

Consider the following quadratic minimax problem: Find z^* such that

$$d^* = \Phi(z^*,i^*) = \min_{z} \max_{i \in I} \Phi(z,i) = \min_{z} \max_{i \in I} \left[a_i + g_i^T z + \tfrac{1}{2} z^T H_i z\right], \quad (13)$$

where I is an index set containing integers $(1,\ldots,r)$ and the matrices H_i, $i = 1,\ldots,r$ are positive definite. Relabeling δx to z and dropping x_k from the notation for $g_i(x_k)$ and $H_i(x_k)$, it is seen that the minimax problem (9) is a special case of the defined quadratic minimax problem.

In order to characterize the solution of the quadratic minimax problem suppose z^* is the solution and suppose $I(z^*)$ is the associated answering index set, containing p integers, relabelling if necessary so that they are the first p positive integer numbers. It follows from the necessary condition (7) that there exist coefficients $c_1^*,\ldots,c_p^*$ such that

$$\sum_{i=1}^{p} c_i^*(g_i + H_i z^*) = 0, \quad (14)$$

with

$$\sum_{i=1}^{p} c_i^* = 1, \quad c_i \geq 0, \; i=1,\ldots,p \quad (15)$$

Moreover, from the definition of the answering set $I(z^*)$, it follows that

$$a_i + g_i^T z^* + \tfrac{1}{2} z^{*T} H_i z^* = \Phi(z^*.i) = d^*, \quad i \in I(z^*). \quad (16)$$

Because c_i^* are non-negative and H_i are positive definite the minimax solution z^*, in terms of the coefficients $c_1^*,\ldots,c_r^*$, is, from (14), given by

$$z^* = -\left(\sum_{i=1}^{p} c_i^* H_i\right)^{-1}\left(\sum_{i=1}^{p} c_i^* g_i\right). \quad (17)$$

Eliminating z^* from (16) by virtue of (17) leads to the following expression for the minimax value d^*, in terms of the coefficients $c_1^*,\ldots,c_p^*$:

$$d^* = \left(\sum_{i=1}^{p} c_i^* a_i\right) - \tfrac{1}{2}\left(\sum_{i=1}^{p} c_i^* g_i\right)^T \left(\sum_{i=1}^{p} c_i^* H_i\right)^{-1} \left(\sum_{i=1}^{p} c_i^* g_i\right) \quad (18)$$

It was originally an intuitive result that coefficients c_i^*, $i=1,\ldots,r$ should be determined by maximizing the expression on the right hand side of (18) with respect to the coefficients c_i, $i=1,\ldots,p$, satisfying the constraint (15). Verification of this intuitive result coupled with consideration of the question which indices form the answering set $I(z^*) \subset I$ has lead to a regularization method with application to a much wider class of problems. The main feature of the regularization method, as applied to the solution of the quadratic minimax problem, is that it transforms the original quadratic minimax problem into a form suitable for the application of the saddle point theorem whereby it becomes possible to interchange the order of the min and max operations. The development is based on the following theorems.

Theorem 3. Let the max function $F_m(z)$ be defined by

$$F_m(z) = \max_{c \in C} \sum_{i=1}^{r} c_i \Phi(z,i) \tag{19}$$

where x $c = (c_1, \ldots, c_r)$, and the domain C is defined by (15). Let the max function $F(z)$ be defined by

$$F(z) = \max_{i \in I} \Phi(z,i). \tag{20}$$

Then, for all bounded z,

$$F_m(z) = F(z), \tag{21}$$

Proof of this and the following theorem is given in the appendix.

Corollary 1. The minimax solution z^* of the modified minimax problem

$$\min_{z} \max_{c \in C} \Phi(z,c) \tag{22}$$

where the modified cost function is given by

$$\Phi(z,c) = \sum_{i=1}^{r} c_i (a_i + g_i^T z + \tfrac{1}{2} z^T H_i z) \tag{23}$$

is equivalent to the minimax solution of the original problem (13).

Theorem 4. The modified cost function $\Phi(z,c)$ possesses a global saddle point, i.e., there exists a point (z^*, c^*) satisfying

$$d^* = \min_{z} \max_{c \in C} \Phi(z,c) = \max_{c \in C} \min_{z} \Phi(z,c). \tag{24}$$

Theorems 3 and 4 allow the solution of the quadratic minimax problem to be obtained by interchanging the order of the min and max operations:

$$d^* = \max_{c \in C} \min_{z} \left[(\sum_{i=1}^{r} c_i a_i) + (\sum_{i=1}^{r} c_i g_i)^T z + \tfrac{1}{2} z^T (\sum_{i=1}^{r} c_i H_i) z \right]. \tag{25}$$

Performing the min operation first gives an analytic expression for the minimizing z in function of c:

$$z^* = -(\sum_{i=1}^{r} c_i H_i)^{-1} (\sum_{i=1}^{r} c_i g_i) \tag{26}$$

and substituting back into (25) gives an analytic expression for the associated min function $G(c)$. The saddle point value d^* is finally obtained by maximizing the min function $G(c)$ inside C:

$$d^* = \max_{c \in C} G(c) = \max_{c \in C} \left[\left(\sum_{i=1}^{r} c_i a_i\right) - \tfrac{1}{2}\left(\sum_{i=1}^{r} c_i g_i\right)^T \left(\sum_{i=1}^{r} c_i H_i\right)^{-1} \left(\sum_{i=1}^{r} c_i g_i\right) \right]. \quad (27)$$

Once the maximizing c^* is obtained from (27) substitution into (25) produces the desired minimax solution z^*. Hence the solution of the quadratic minimax problem has been reduced to a maximization under linear constraints.

Comparison of (27) with (18) fully justifies the original intuitive result. Moreover, it is seen from (27) that all coefficients c_i^*, $i=1,\ldots,r$ are determined from the maximization of the min function while the derivation of this result from the necessary condition (7) implies that optimal values c_j^*, corresponding to indices $j \in I$ which do not enter into the answering set $I(x^*)$ are equal to zero.

NEWTON-RAPHSON ALGORITHM FOR THE CONVEX MINIMAX PROBLEM

Application of the results derived in the previous section allows the construction of the Newton-Raphson algorithm in which the solution of the convex minimax problem is obtained by a succession of two related but numerically decoupled maximization problems:

(a) Initialize the procedure by selecting x_o and place $k = 0$;

(b) Given x_k find the answering set $Y(x_k)$ by solving the first maximization problem:

$$F(x_k) = \max_{y \in Y} J(x_k, y);$$

(c) Given the finite answering set $Y(x_k)$ compute $g_i(x_k)$ and $H_i(x_k)$, $y_i \in Y(x_k)$, and find c_i^*, $i=1,\ldots,r$ by solving the second maximization problem:

$$d_k = \max_{c \in C} -\left[\tfrac{1}{2}\left(\sum_{i=1}^{r} c_i g_i\right)^T \left(\sum_{i=1}^{r} c_i H_i\right)^{-1} \left(\sum_{i=1}^{r} c_i g_i\right)\right]$$

where the domain C is defined by (14);

(d) Given c_i^*, $i=1,\ldots,r$ find the optimal step size

$$\delta x^* = -\left(\sum_{i=1}^{r} c_i^* H_i\right)^{-1} \left(\sum_{i=1}^{r} c_i^* g_i\right)$$

and take as the next nominal point

$$x_{k+1} = x_k + \delta x_k^*;$$

(e) Place $k = k+1$ and repeat from step (b) if the step size is not sufficiently small.

The described Newton-Raphson algorithm solves the convex minimax problem in principle only. Modifications of the algorithm are necessary to assure convergence, i.e., to avoid oscillations which are a consequence of the upper semi-continuity of the answering set $Y(x)$ as a function of x. Reconsidering the partial expansion formula (9) (a fortiori the full expansion formula of Demjanov) it may be deduced that the expansion is valid only within the sphere with center at x_k and radius equal to the distance from x_k to the nearest to it (but not coinciding with it) point of semi-continuity of the answering set $Y(x)$. As in the proposed gradient type algorithms [2,4] the basic remedy against the semi-continuity of the answering set $Y(x)$ is to use in the algorithm not the answering set $Y(x)$ but an extended answering set $Y(x,e)$ [4]:

Definition 3. Given $\underline{x} \in X$, the extended answering set is denoted by $Y(\underline{x},e)$ where e is a given positive constant, and is defined by

$$Y(\underline{x},e) = \left\{ \underline{y} : \underline{y} \in Y,\ F(\underline{x}) - J(\underline{x},\underline{y}) \leq e,\ \underline{y} \text{ is a local maximum} \right\}.$$

Convergence of the Newton-Raphson algorithm is now based on the fact that for sufficiently large $e > 0$, the extended answering set $Y(x,e)$ is continuous with respect to x [4]. On the other hand selection of a reasonably small value for e is based on the assumption that local maxima sufficiently far below the global maximum will not affect convergence. However, employing the extended answering set $Y(x_k,e)$ in each iteration, points $y_i \in Y(x_k,e)$ for which $J(x_k,y_i) < F(x_k)$ appear and while the expansion formula (9) may still be employed it is more efficient to return to the original definition of the max function and the expansion formula for $J(x,y_i)$ and write:

$$F(x) = \max_{y_i \in Y(x_k,e)} \left[J(x_k,y_i) + g_i(x_k)^T \delta x + \tfrac{1}{2} \delta x^T H_i(x_k) \delta x \right]. \qquad (28)$$

Evidently, the problem of selecting the optimal step size δx^* is now equivalent to the solution of the general quadratic minimax problem (13) and the modified Newton-Raphson algorithm consists in the following:

(a) Initialize the procedure by selecting x_o and place $k = 0$;

(b) Given x_k find the extended answering set $Y(x_k,e)$ by solving the first maximization problem:

$$F(x_k) = \max_{y \in Y} J(x_k,y); \qquad (29)$$

(c) Given the finite answering set $Y(x_k,e)$, compute $a_i = J(x_k,y_i)$, $g_i(x_k)$ and $H_i(x_k)$, $y_i \in Y(x_k,e)$, and find c_i^*, $i=1,\ldots,R$ by solving the second maximization problem:

$$d_k = \max_{c \in C} \left[\left(\sum_{i=1}^{R} c_i a_i\right) - \tfrac{1}{2}\left(\sum_{i=1}^{R} c_i g_i\right)^T \left(\sum_{i=1}^{R} c_i H_i\right)^{-1} \left(\sum_{i=1}^{R} c_i g_i\right) \right]; \qquad (30)$$

(d) Given c_i^*, $i=1,\ldots,R$ find the optimal step size

$$x^* = -\left(\sum_{i=1}^{R} c_i^* H_i\right)^{-1} \left(\sum_{i=1}^{R} c_i^* g_i\right)$$

and take as the next nominal point

$$x_{k+1} = x_k + \delta x^*;$$

(e) Place $k = k+1$ and repeat from step (b) if the step size is not sufficiently small.

NUMERICAL ASPECTS

The two maximizations that have to be performed in each iteration of the Newton-Raphson algorithm have completely different computational aspects. The difficulty of the first maximization problem depends on the form of the cost function J(x,y) and the domain Y. This problem is standard in all minimax algorithms and was here treated by the same maximization methods. In particular, if the dimensionality of y is rather small the definition of a grid of points over Y and maximization over the grid rather than over the whole Y has been found quite efficient and sufficiently accurate in some gradient and elimination type algorithms. Another efficient procedure was found to be Rosenbrock's rotating coordinate search.

The second maximization problem is a unique feature of the proposed Newton-Raphson algorithm and replaces the search for the negative gradient direction and the one dimensional minimization of the max function along the negative gradient direction in gradient type algorithms, or, the reduction of the union of answering sets and the solution of the associated minimax problem in Salmon's exchange type algorithm[5]. This maximization problem has been found particularly simple due to the fact that the resulting min function G(c), given in (27), is concave in C and that C is a convex set defined by a set of linear constraints. These features make the problem particularly suitable for the application of the gradient projection algorithm which has been found completely satisfactory in the solution of this maximization problem.

CONCLUSIONS

In this paper the complete and original solution of the quadratic minimax problem has been presented. The results have been applied in the solution of the convex minimax problem through the construction of a Newton-Raphson type minimax algorithm. The results are based on a regularization of the quadratic minimax problem which consists in the use of a modified cost function which, on the one side, leaves the max function unchanged, and on the other allows the application of the saddle point theorem and the interchange of the order of the min and max operations. Computational aspects, speed of convergence and comparison with other proposed minimax algorithms shall have to be considered in future papers. Computational experience gained so far has been promising. In particular, the solution of the quadratic minimax problem,(which finds only one application in the Newton-Raphson algorithm) by maximizing the min function G(t) by the gradient projection algorithm is very efficient. Problems where the dimensionality of z is five and the index set I contains twenty elements have been solved with ease. The efficiency of the Newton-Raphson algorithm itself does depend on the judicious choice of the extended answering set $Y(x_k, e)$ but this was the only problem encountered and was always resolved after a few trial runs.

ACKNOWLEDGEMENT

Investigations reported in this paper were initialized while the author was a visiting research associate with the Coordinated Science Laboratory, University of Illinois and were supported in part by the NSF Grant GK-3893. The work was continued at the "Mihailo Pupin" Institute under a grant from the National Scientific Council.

REFERENCES

1. V.F. Demjanov, "On the solution of some minimax problems," Part I, Kibernetika, No. 6, 1966, pp.58-66; Part II, Kibernetika, No. 3, 1967, pp.62-66.

2. V.F. Demjanov, "Algorithms for some minimax problems," J. Comp. Sys. Sci., No. 2, 1968, pp.342-380.

3. V.F. Demjanov, "Differentiation of the maxmin function," Parts I, II and III, J. Computational Math. and Math. Phys., Vol. 10, 1970 (in russian).

4. J.E. Heller, "A gradient algorithm for minimax design," Ph.D. Thesis Report R-406, Coordinated Science Lab., University of Illinois, Jan., 1969.

5. D. Salmon, "Minimax Controller Design," IEEE Trans. on Automatic Control, Vol. AC-13, No. 3, July, 1968, pp.369-376.

6. M.R. Osborn and G.A. Watson, "On the best linear Chebyshev approximation," Computer J., Vol. 10, 1967, pp.172-177.

7. M.R. Osborn and G.A. Watson, "An algorithm for minimax approximation in the nonlinear case," Computer J., Vol. 12, 1969, pp.63-68.

8. G.A. Watson, "On an algorithm for nonlinear minimax approximation," Communications of the ACM, Vol. 13, No. 3, March, 1970, pp.160-162.

9. J. Medanic, "Elimination algorithm for computation of the minimax," Proc. IV Allerton Conf. on Circuit and System Theory, University of Illinois, Urbana, Ill., Oct., 1968.

10. V. Ciric, "Design of Minimum sensitivity control systems," Ph.D. Thesis, Rice University, Houston, Texas, 1969.

11. J. Medanic, "Theoretical and Computational aspects of the minimax problem," Proc. XIV ETAN Conf., Sarajevo, Yugoslavia, June, 1970 (in serbo-croatian).

12. Fan, K., "Minimax theorems," Proc. Natl. Acad. Sci., Vol. 39, 1953.

13. S. Kakutani, "Generalization of Brouwer's fixed point theorem," Duke Math. J., Vol. 8, No. 3, 1941.

14. J.M. Danskin, The theory of Max-Min, Series Econometrics and Operations Research, Vol. V, Springer-Verlag, New York, 1967.

APPENDIX

To prove Theorem 3, let $c^* \in C$ maximize in (19) and let $\underline{c} \in C$ be such that

$$\sum_{i=1}^{r} \underline{c}_i \Phi(z,i) = \max_{i \in I} \Phi(z,i) = F(z). \tag{A-1}$$

Clearly,

$$F_m(z) = \sum_{i=1}^{r} c_i^{*}\Phi(z,i) \geq \sum_{i=1}^{r} \underline{c}_i\Phi(z,i) = \max_{i \in I} \Phi(z,i) = F(z). \qquad \text{(A-2)}$$

On the other hand, since $\Phi(z,i) \leq F(z)$, $i \in I$, it follows that

$$\sum_{i=1}^{r} c_i^{*}\Phi(z,i) \leq \sum_{i=1}^{r} c_i^{*}F(z) \qquad \text{(A-3)}$$

and therefore

$$F_m(z) = \sum_{i=1}^{r} c_i^{*}\Phi(z,i) \leq F(z) \qquad \text{(A-4)}$$

which together with (A-2) proves equivalence of $F_m(z)$ and $F(z)$.

To prove Thoerem 4 it is necessary to show that the domains of z and c are convex and that the modified cost function (23) is convex in z and concave in c. The domain of z is not restricted and hence is convex, while the domain C is convex since it is defined as the intersection of two convex sets: the positive hyperquadrant $c_i \geq 0$, $i=1,\dots,r$ and the hyperplane

$$\sum_{i=1}^{r} c_i = 1. \qquad \text{(A-5)}$$

Moreover, since the c_i are non-negative and the matrices H_i are positive definite, the modified cost function remains convex with respect to z, and because the modified cost function is linear in c_i, $i=1,\dots,r$, it is also concave in c, which completes the proof.

CONDITIONS FOR A MINIMAX OPTIMUM

JOHN W. BANDLER
Department of Electrical Engineering
McMaster University
Hamilton, Ontario, Canada

Abstract This paper derives and discusses necessary conditions for an optimum in nonlinear minimax approximation problems. A straightforward geometrical interpretation is presented. The results may be used to test for convergence in computer-aided network optimization, in tests for optimality in the Chebyshev sense of any given design and to gain insight which may be helpful in developing minimax approximation algorithms.

INTRODUCTION

Of considerable importance to network designers, whether they use synthesis methods or optimization methods, is the optimality of their design. Conditions for optimality in the Chebyshev sense in conventional synthesis problems involving polynomials and rational functions are fairly widely appreciated. However, with the ever increasing need for designs containing elements not conducive to the rational function approach, e.g., a mixture of lumped and distributed elements, and the inevitable application of automatic optimization methods involving least pth [1-3] and minimax [4-10] objectives, some means of testing for convergence to an optimum for more arbitrary problems is highly desirable.

In this paper the necessary conditions for a minimax optimum for a general nonlinear minimax approximation problem are obtained from the Kuhn-Tucker necessary conditions for a constrained optimum in nonlinear programming [11]. Just as the Kuhn-Tucker relations have a straightforward geometrical interpretation, so the conditions for a minimax optimum also have such an interpretation. This interpretation is presented and discussed. The restrictions for which the necessary conditions are sufficient are also discussed.

A comparison with theorems characterizing minimax approximations presented by Curtis and Powell [12] is made. Finally, illustrative examples demonstrate the implementation of the necessary conditions.

THE PROBLEM

Let us define errors related to the "upper" specifications and "lower" specifications, respectively, as [8]

$$\begin{aligned} e_u(\underline{\phi},\psi) &\triangleq w_u(\psi)\ (F(\underline{\phi},\psi) - S_u(\psi)) \\ e_\ell(\underline{\phi},\psi) &\triangleq w_\ell(\psi)\ (F(\underline{\phi},\psi) - S_\ell(\psi)) \end{aligned} \qquad (1)$$

where

$F(\underline{\phi},\psi)$	is the approximating function (network response)
$S_u(\psi)$	is an upper specified function (desired response bound)
$S_\ell(\psi)$	is a lower specified function (desired response bound)
$w_u(\psi)$	is an upper weighting function
$w_\ell(\psi)$	is a lower weighting function

This work was supported by the McMaster University Science and Engineering Division Research Board.

$\underline{\phi}$ represents the k independent parameters
ψ is an independent variable (e.g., frequency or time).

We will make a number of assumptions which are usually fulfilled in practice. The functions $e_u(\underline{\phi},\psi)$ and $e_\ell(\underline{\phi},\psi)$ are evaluated at a finite discrete set of values of ψ taken from one or more closed intervals over each of which the functions are continuous in ψ. Therefore, let

$$\begin{aligned} e_{ui}(\underline{\phi}) &\triangleq e_u(\underline{\phi}, \psi_i) && i \in I_u \\ e_{\ell i}(\underline{\phi}) &\triangleq e_\ell(\underline{\phi}, \psi_i) && i \in I_\ell \end{aligned} \tag{2}$$

where I_u and I_ℓ are appropriate index sets.

The discrete approximation problem can be written in the form:

$$\text{minimize } U = \phi_{k+1} \tag{3}$$

subject to

$$\begin{aligned} \phi_{k+1} - e_{ui}(\underline{\phi}) &\geq 0 && i \in I_u \\ \phi_{k+1} + e_{\ell i}(\underline{\phi}) &\geq 0 && i \in I_\ell \end{aligned} \tag{4}$$

This is equivalent to the problem of finding

$$\min_{\underline{\phi}}\{\max_{i}[e_{ui}(\underline{\phi}), -e_{\ell i}(\underline{\phi})]\} . \tag{5}$$

It is assumed that a minimum exists in a closed and bounded region of points $\underline{\phi}$ and that $e_{ui}(\underline{\phi})$ and $e_{\ell i}(\underline{\phi})$ are continuous for all i with continuous partial derivatives at least in the neighborhood of the minimum.

THE CONDITIONS

Theorem 1 At an optimum point $\underline{\phi}^o$ for the minimax approximation problem

$$\sum_{i\varepsilon J_u} u_{ui}\, \underline{\nabla}\, e_{ui}(\underline{\phi}^o) = \sum_{i\varepsilon J_\ell} u_{\ell i}\, \underline{\nabla}\, e_{\ell i}(\underline{\phi}^o)$$

$$\sum_{i\varepsilon J_u} u_{ui} + \sum_{i\varepsilon J_\ell} u_{\ell i} = 1$$

$$u_{ui} \geq 0 \qquad i \in J_u$$

$$u_{\ell i} \geq 0 \qquad i \in J_\ell$$

where $e_{ui}(\underline{\phi}^o)$ for $i \in J_u$ and $-e_{\ell i}(\underline{\phi}^o)$ for $i \in J_\ell$ are the equal maxima. The special case $S_{ui} = S_{\ell i} = S_i$, $w_{ui} = w_{\ell i} = w_i$, $I_u = I_\ell = I$ is also accommodated.

Proof Invoking the Kuhn-Tucker relations, we may write at $\underline{\phi} = \underline{\phi}^o$

$$\begin{bmatrix} \underline{\nabla} U \\ \dfrac{\partial U}{\partial \phi_{k+1}} \end{bmatrix} = \sum_{i\varepsilon I_u} u_{ui} \begin{bmatrix} \underline{\nabla} \\ \dfrac{\partial}{\partial \phi_{k+1}} \end{bmatrix} (\phi_{k+1} - e_{ui}(\underline{\phi})) + \sum_{i\varepsilon I_\ell} u_{\ell i} \begin{bmatrix} \underline{\nabla} \\ \dfrac{\partial}{\partial \phi_{k+1}} \end{bmatrix} (\phi_{k+1} + e_{\ell i}(\underline{\phi})) \tag{6}$$

$$u_{ui}(\phi_{k+1} - e_{ui}(\underline{\phi})) = 0 \qquad i \in I_u$$
$$u_{\ell i}(\phi_{k+1} + e_{\ell i}(\underline{\phi})) = 0 \qquad i \in I_\ell \tag{7}$$

where

$$u_{ui} \geq 0 \qquad i \in I_u$$
$$u_{\ell i} \geq 0 \qquad i \in I_\ell \tag{8}$$

At a minimum at least one of the constraints (4) must equal zero or U could be further reduced without violation of the constraints. Let J_u be an index set corresponding to a possible n_u constraints related to the "upper" specifications which are zero and J_ℓ be an index set corresponding to a possible n_ℓ constraints related to the "lower" specifications which are zero. Then

$$u_{ui} \begin{cases} \geq 0 & i \in J_u \\ = 0 & i \notin J_u \end{cases}$$
$$u_{\ell i} \begin{cases} \geq 0 & i \in J_\ell \\ = 0 & i \notin J_\ell \end{cases} \tag{9}$$

Note also that

$$\underline{\nabla} U = \underline{\nabla}\phi_{k+1} = \underline{0}$$
$$\frac{\partial U}{\partial \phi_{k+1}} = 1 \tag{10}$$
$$\frac{\partial e_{ui}(\underline{\phi})}{\partial \phi_{k+1}} = \frac{\partial e_{\ell i}(\underline{\phi})}{\partial \phi_{k+1}} = 0$$

everywhere. The necessary conditions now become:

$$\begin{bmatrix} \underline{0} \\ 1 \end{bmatrix} = \sum_{i \in J_u} u_{ui} \begin{bmatrix} -\underline{\nabla} e_{ui}(\underline{\phi}^o) \\ 1 \end{bmatrix} + \sum_{i \in J_\ell} u_{\ell i} \begin{bmatrix} \underline{\nabla} e_{\ell i}(\underline{\phi}^o) \\ 1 \end{bmatrix} \tag{11}$$

from which the theorem follows.

Geometrically, the necessary conditions have a straightforward interpretation. To see this let the problem be written in the form:

$$\text{minimize } U = \phi_{k+1} \tag{12}$$

subject to

$$\phi_{k+1} \geq f_i(\underline{\phi}) \qquad i = 1, 2, \ldots, m \tag{13}$$

The necessary conditions for an optimum are

$$\sum_{i=1}^{m_o} u_i \, \underline{\nabla}\, f_i(\underline{\phi}^o) = \underline{0} \tag{14}$$

$$\sum_{i=1}^{m_o} u_i = 1 \qquad (15)$$

$$u_i \geq 0 \qquad\qquad i = 1, 2, \ldots, m_o \qquad (16)$$

where it is assumed, for convenience, that the first m_o constraints are equalities or that the first m_o of f_i are equal maxima.

Figure 1 shows examples for k=2. Observe that the conditions are not satisfied in Figures 1(a) and (b). They are satisfied in Figures 1(c), (d) and (e). The points indicated in Figures 1(d) and (e) are clearly optimum points.

Theorem 2 If the relations in Theorem 1 are satisfied at a point $\underline{\phi}^o$ and the $e_{ui}(\underline{\phi})$ for $i \in I_u$ are convex and the $e_{\ell i}(\underline{\phi})$ for $i \in I_\ell$ are concave then $\underline{\phi}^o$ is optimal.

The proof for the above sufficiency theorem would follow similar lines to those required for proving the sufficiency of the Kuhn-Tucker relations.

Now let

$$\underline{D} \triangleq [\underline{d}_1\ \underline{d}_2\ \ldots\ \underline{d}_r] \qquad (17)$$

and

$$\underline{\rho} \triangleq [\rho_1\ \rho_2\ \ldots\ \rho_r]^T \qquad (18)$$

where

$$r = n_u + n_\ell \qquad (19)$$

and where each $\underline{d}_j$ or ρ_j for $j=1,2,\ldots,n_u$ corresponds to a $\underline{\nabla} e_{ui}(\underline{\phi}^o)$ or u_{ui}, respectively, for $i \in J_u$ and each $\underline{d}_j$ or ρ_j for $j=n_u+1, n_u+2, \ldots, r$ corresponds to a $\underline{\nabla} e_{\ell i}(\underline{\phi}^o)$ or $-u_{\ell i}$, respectively, for $i \in J_\ell$. Then at an optimum, from Theorem 1

$$\underline{D}\ \underline{\rho} = \underline{0} \qquad (20)$$

is satisfied for at least one $\rho_j \neq 0$.

The two theorems of Curtis and Powell [12] concerning minimax approximations follow immediately. Adapting them to the present notation, the first one would state that the rank of $\underline{D}$ is less than r; the second that, for r=k+1, the signs of the maximum deviations are either all the same as or all opposite to the signs of the corresponding elements of $\underline{\rho}$. The present Theorem 1 appears essentially to contain both of the theorems of Curtis and Powell and without emphasizing the special case r=k+1.

EXAMPLES

Curtis and Powell [12] gave two examples to illustrate the application of their theorems. The necessary conditions derived here confirm their results.

The necessary conditions will be applied to verify the optimality of two optimum quarter-wave transmission-line transformers. The examples were previously studied by Bandler and Macdonald [7],[9]. The necessary gradient vectors for both examples are derived by Bandler and Seviora in another paper presented at this conference using the adjoint network approach [13]. Tables I and II summarize the problems and list the partial derivatives.

By inspection, multipliers in the ratio

3 : 2 : 1

for Table I, and by solving three equations in three unknowns

3.000 : 2.812 : 1.760 : 1.000

for Table II can be found (within rounding errors) for the gradient vectors to satisfy the necessary conditions for optimality.

It is also interesting to relate the gradient vectors in Table I to the various contour diagrams given by Bandler and Macdonald [7]. If only Z_1 and Z_2 are assumed variable, then the design may be optimal over the band 0.5 to 1.0GHz as well as 0.5 to 1.5 GHz. If only ℓ_1 and ℓ_2 are assumed variable, the design may be optimal over the band 0.5 to 1.5 GHz but not from 0.5 to 1.0 GHz. These and other conclusions can be quickly reached by applying Theorem 1 as easily as from physical considerations. The three section design (Table II) can be investigated in a similar manner, if desired.

CONCLUSIONS

Observe that no assumptions on the number of equal extrema as compared with the number of parameters has been made. Further, note that if (by optimization or otherwise)we have an almost equal-ripple solution, the conditions might still be valid if it is assumed that the weighting functions are readjusted to equalize the extrema. Consider, for example, Figure 1(d). Thus the progress of least pth approximation [1-3] could also be checked.

Very wide application of these ideas to approximation problems in network design is envisaged, particularly when nonclassical procedures have to be employed. The capability of evaluating gradient vectors associated with network responses is required. Using the adjoint network method [3], [13], [14], this can be readily done for a wide class of networks.

ACKNOWLEDGEMENT

The author would like to acknowledge numerous and useful discussions on the subject matter of this paper with Dr. E. Della Torre of the Department of Electrical Engineering, McMaster University, and with R. E. Seviora of the Department of Electrical Engineering, University of Toronto. The programming assistance of A. Lee-Chan of the McMaster Data Processing and Computing Centre is acknowledged.

REFERENCES

[1] G.C. Temes and D.Y.F. Zai, "Least pth approximation", IEEE Trans. Circuit Theory, vol. CT-16, pp. 235-237, May 1969.

[2] G.C. Temes, "Optimization methods in circuit design", in Computer Oriented Circuit Design, F.F. Kuo and W.G. Magnuson, Jr., Eds. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

[3] J.W. Bandler and R.E. Seviora, "Current trends in network optimization", IEEE Trans. Microwave Theory and Techniques, vol. MTT-18, December 1970.

[4] A.D. Waren, L.S. Lasdon and D.F. Suchman, "Optimization in engineering design", Proc. IEEE, vol. 55, pp. 1885-1897, November 1967.

[5] Y. Ishizaki and H. Watanabe, "An iterative Chebyshev approximation method for network design", IEEE Trans. Circuit Theory, vol. CT-15, pp. 326-336, December 1968.

[6] M.R. Osborne and G.A. Watson, "An algorithm for minimax approximation in the nonlinear case", Computer J., vol. 12, pp. 63-68, February 1969.

[7] J.W. Bandler and P.A. Macdonald, "Cascaded noncommensurate transmission-line networks as optimization problems", IEEE Trans. Circuit Theory, vol. CT-16, pp. 391-394, August 1969.

[8] J.W. Bandler, "Optimization methods for computer-aided design", IEEE Trans. Microwave Theory and Techniques, vol. MTT-17, pp. 533-552, August 1969.

[9] J.W. Bandler and P.A. Macdonald, "Optimization of microwave networks by razor search", IEEE Trans. Microwave Theory and Techniques, vol. MTT-17, pp. 552-562, August 1969.

[10] J.W. Bandler, "Computer optimization of inhomogeneous waveguide transformers", IEEE Trans. Microwave Theory and Techniques, vol. MTT-17, pp. 563-571, August 1969.

[11] H.W. Kuhn and A.W. Tucker, "Non-linear programming", Proc. 2nd Symp. on Math. Statistics and Probability. Berkeley, Calif.: University of California Press, 1951, pp. 481-493.

[12] A.R. Curtis and M.J.D. Powell, "Necessary conditions for a minimax approximation", Computer J., vol. 8, pp. 358-361, January 1966.

[13] J.W. Bandler and R.E. Seviora, "Sensitivities in terms of wave variables", 8th Annual Allerton Conference on Circuit and System Theory, Urbana, Illinois, October 1970.

[14] S.W. Director and R.A. Rohrer, "The generalized adjoint network and network sensitivities", IEEE Trans. Circuit Theory, vol. CT-16, pp. 318-323, August 1969.

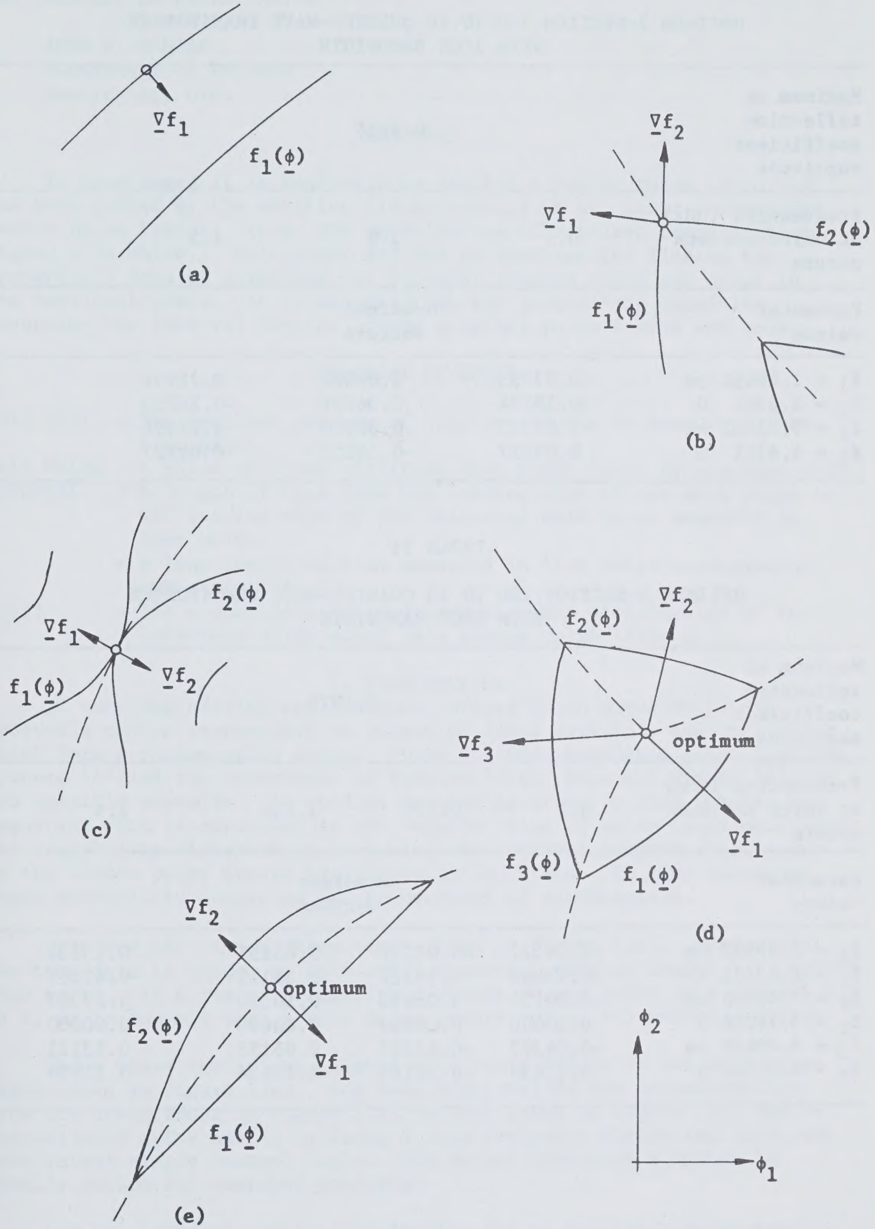

$\underline{\nabla f}_1$
$f_1(\underline{\phi})$
(a)
$\underline{\nabla f}_2$
$\underline{\nabla f}_1$
$f_2(\underline{\phi})$
$f_1(\underline{\phi})$
(b)
$f_2(\underline{\phi})$
$\underline{\nabla f}_1$
$\underline{\nabla f}_2$
$f_1(\underline{\phi})$
(c)
$\underline{\nabla f}_2$
$f_2(\underline{\phi})$
$\underline{\nabla f}_3$
optimum
$\underline{\nabla f}_1$
$f_3(\underline{\phi})$
$f_1(\underline{\phi})$
(d)
$\underline{\nabla f}_2$
optimum
$f_2(\underline{\phi})$
$\underline{\nabla f}_1$
$f_1(\underline{\phi})$
(e)
ϕ_2
ϕ_1

Figure 1. Geometrical interpretation. The necessary conditions are not satisfied in (a) or (b). They are satisfied in (c), (d) and (e).

TABLE I

OPTIMUM 2-SECTION 10Ω TO 1Ω QUARTER-WAVE TRANSFORMER WITH 100% BANDWIDTH

Maximum of reflection coefficient magnitude		0.4286	
Frequencies (GHz) at which maximum occurs	0.5	1.0	1.5
Parameter values		Gradient vectors	
$\ell_1 = 7.49482$ cm	-0.07333	0.00000	0.21998
$Z_1 = 2.2361\ \Omega$	-0.18254	0.36506	-0.18253
$\ell_2 = 7.49482$ cm	-0.07333	0.00000	0.21998
$Z_2 = 4.4721\ \Omega$	0.09127	-0.18253	0.09127

TABLE II

OPTIMUM 3-SECTION 10Ω TO 1Ω QUARTER-WAVE TRANSFORMER WITH 100% BANDWIDTH

Maximum of reflection coefficient magnitude			0.1973	
Frequencies (GHz) at which maximum occurs	0.5	0.770	1.230	1.5
Parameter values			Gradient vectors	
$\ell_1 = 7.49482$ cm	-0.04377	-0.03227	0.05155	0.13131
$Z_1 = 1.63471\ \Omega$	-0.43556	0.38111	0.38111	-0.43555
$\ell_2 = 7.49482$ cm	-0.09129	0.06453	-0.10310	0.27387
$Z_2 = 3.16228\ \Omega$	0.00000	0.00000	0.00000	0.00000
$\ell_3 = 7.49482$ cm	-0.04377	-0.03227	0.05155	0.13131
$Z_3 = 6.11729\ \Omega$	0.11639	-0.10184	-0.10184	0.11639

AN ANALYTICAL METHOD FOR FINDING THE INTERVAL PROBABILITY DENSITY OF TWO INTERLEAVED PULSE TRAINS

JOHN H. O'HARA
Department of Defense
Washington, D.C.

ABSTRACT

In many cases it is desirable to analyze a random pulse train that has been formed by the addition (interleaving) of two other independent random pulse trains. (e.g. the interleaving of a pulsed communication signal with noise.) This paper derives an equation for finding the probability density governing the interval lengths which can occur in the resultant train. It is assumed that the probability densities governing the interval lengths of the original pulse trains are known.

GLOSSARY OF TERMS

TIME UNIT	- The smallest measure of time one wishes to measure on the time axis.
UNIT PULSE	- A pulse of fixed amplitude with width equal to one time unit.
INTERVAL	- A length of time from the leading edge of one unit pulse to the leading edge of the following unit pulse measured in time units.
T	- A long length of time measured in time units encompassing many intervals.
$P_q(I)$	- The probability density governing the distribution of the intervals which occur in a random pulse train Q.

I. INTRODUCTION

In many engineering applications, events which occur at random intervals can be represented as pulses of fixed amplitude and unit width which form a random pulse train. Pulse Position Modulated communications systems (1) and the occurrence of Intracellulary Neuronal Events (2) are two specific examples. In studies related to areas of this nature, the important data is contained in the relative time of pulse occurrence hence the probability distribution governing the interval lengths which occur in the random pulse trains are extremely important. In many instances these probability densities can be measured or approximated.

Consider the random pulse train shown in figure (1a). Assume that the time axis is broken up into small segments, each of length Δt, called time units. If a positive transition occurs at any point in a segment Δt, it is assumed that a unit pulse occurs during the entire segment.

Under these conditions, figure (1a) would reduce to the unit pulse train shown in figure (1b). For many applications the transformation from the train shown in figure (1a) to that shown in figure (1b) can be accomplished quite easily by using a high frequency digitizing clock and some rather simple control logic. The output from such a system is ideally suited for computer analysis.

Let the interval probability density for an arbitrary pulse train Q be defined as

$$P_q(i) = P(q = i)$$

This defines the probability that an arbitrary interval which occurs in a pulse train Q, will have a length equal to i time units. Since the time axis has been quantized into time segments of unit length, the interval probability densities will be discrete. All possible interval lengths from one time unit up to q_{max} time units are allowed. q_{max} is the maximum interval length which can occur in the pulse train Q.

Quite frequently one may encounter a pulse train that has been formed by the addition of two independent pulse trains. A typical example is a pulsed communications signal occurring in an anomalous type noise. The statistical characteristics governing the interval distribution for individual pulse trains (i.e. the signal train and noise train) may be known. It is desired to find the statistics governing the interval distribution of the summed train (i.e. the signal plus noise train) based on the interval statistics of the original trains. Figure 2.

Consider the two independent, random, stationary pulse trains X, Y given in figure 3. Here the pulses are represented as lines for ease of illustration. It is assumed that the interval probability density of each train is known. Call these densities $P_x(k)$ and $P_y(m)$. Let E(X) be the expected interval length in the X train and E(Y) be the expected interval length in the Y train. E(X) and E(Y) can be found from $P_x(k)$ and $P_y(m)$

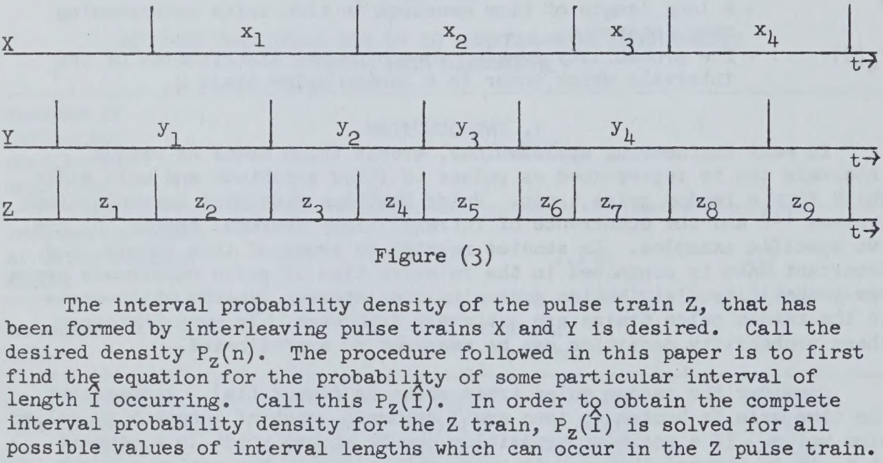


Figure (3)

The interval probability density of the pulse train Z, that has been formed by interleaving pulse trains X and Y is desired. Call the desired density $P_z(n)$. The procedure followed in this paper is to first find the equation for the probability of some particular interval of length $\hat{I}$ occurring. Call this $P_z(\hat{I})$. In order to obtain the complete interval probability density for the Z train, $P_z(\hat{I})$ is solved for all possible values of interval lengths which can occur in the Z pulse train.

II. ANALYSIS

For derivation purposes, it is advantageous to consider this problem in two parts. Part 1 accounts for intervals that begin on pulses contributed by the Y pulse train. Part 2 accounts for intervals that begin on pulses contributed by the X pulse train.

Part 1

In Part 1 it is desired to find the probability that an interval of length $(\hat{I})$ occurred in the Z pulse train, given that the pulse which started the interval, occurred in the Y pulse train. Call this probability $P_z(\hat{I}/Y)$.

Take any segment G of the Z train, which starts on a pulse B_1 contributed by the Y train, and which ends on the next pulse A_1 contributed by the X train. See figure 4.

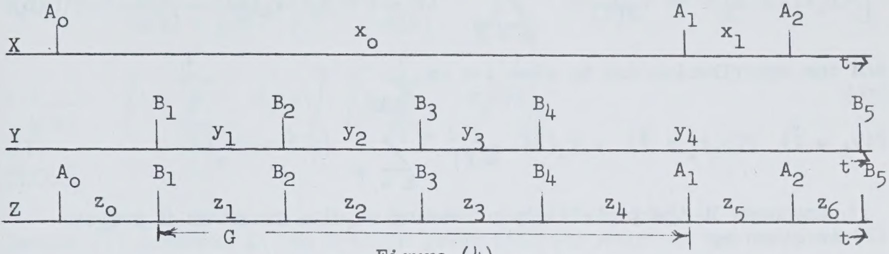


Figure (4)

Observation of the segment G reveals that intervals in the Z train may be formed in two ways when the interval starts with a B pulse.*

1a. Intervals may occur due to the Y pulse train only.

(e.g., z_1, z_2, z_3) if $B_iA_1 \geq y_i$

1b. Intervals may occur due to a B pulse from the Y train combining with an A pulse from the X train (e.g. z_4) if $B_iA_1 \leq y_i$

Hence $P_z(\hat{I}/Y) = P(\text{case 1a}) + P(\text{case 1b})$

For case 1a, the probability of having an interval of length $(\hat{I})$ can be given by

$$P(y_i = \hat{I})\ P(B_iA_1 \geq \hat{I}) \tag{1}$$

$P(y_i = \hat{I})$ can be found by letting $m = \hat{I}$ in the known density function $P_y(m)$, and $P(B_iA_1 \geq \hat{I})$ can be found in the following manner.

Since it is assumed that trains X and Y are completely independent, one must resort to an averaging process over all possible ways in which $B_iA_1 \geq \hat{I}$. Let A_o be the last pulse in the X train preceding the occurrence of B_i in the Y train. Let j be the length of the interval formed for an arbitrary occurrence of A_o and B_i. Then

$$\left[P(B_iA_1) \geq \hat{I}\right] = \sum_{j=0}^{x_{max}-\hat{I}} P(A_oB_i = j)\ P(A_oA_1 \geq \hat{I} + j) \tag{2}$$

Now $P(A_oB_i = j) = \dfrac{1}{E(X)}$ **(3)

and
$$P(A_oA_1 \geq \hat{I} + j) = \sum_{k=\hat{I}+j}^{x_{max}} P_x(k) \tag{4}$$

hence (2) becomes

$$\left[P(B_iA_1) \geq \hat{I}\right] = \frac{1}{E(X)} \sum_{j=0}^{x_{max}-\hat{I}} \ \sum_{k=\hat{I}+j}^{x_{max}} P_x(k) \tag{5}$$

---

*If an A pulse and a B pulse occur simultaneously, the event will be considered as a B pulse occurrence in Part 1.

**The author has found equation 3 to be a bone of contention from time to time. He will gladly send a proof of this equation to anyone upon request.

The lower limit on k is $\hat{I}$ and the upper limit is x_{max} hence (5) becomes

$$\left[P(B_iA_1) \geq \hat{I}\right] = \frac{1}{E(X)} \sum_{k=\hat{I}}^{x_{max}} (k - \hat{I} + 1)\ P_x(k) \qquad (6)$$

and the contribution due to case 1-a is

$$P(y_i = \hat{I})\ \ P(B_iA_1 \geq \hat{I}) = P_y(\hat{I})\ \frac{1}{E(X)} \sum_{k=\hat{I}}^{x_{max}} (k-\hat{I}+1)P_x(k) \qquad (7)$$

For case 1b the probability of having a pulse interval of length $(\hat{I})$ is given by

$\left[P(y_i > \hat{I})\right]$ $\left[P(\text{Pulse in X train, call it } A_1\text{, occurs precisely at}\right.$
$\left.B_i + \hat{I})\right]$ $\left[P(\text{no other X train pulses occur in the interval } B_iA_1)\right]$ (8)

Now

$$P(y_i > \hat{I}) = 1-P(Y_i \leq \hat{I}) = I - \sum_{m=I}^{\hat{I}} P_y(m) \qquad (9)$$

Note that using the limit $\hat{I}$ eliminates the case when $y_i = \hat{I}$. This case has been accounted for in Part 1a. Since $P_y(m)$ is given, (9) can be obtained.

Now

$P(\text{Pulse in X train occurs precisely at } (B_i + \hat{I})) = I/E(X)$ (10)

and

$P(\text{no other X train pulses occur in } B_i\ A_1)$ is $\quad 1- \sum_{k=1}^{\hat{I}} P_x(k)$ (11)

The limit $\hat{I}$ in (11) eliminates the probability of counting the case when A_o and B_i occur simultaneously and $x_o = B_iA_1 = \hat{I}$. Since $P_x(k)$ is given, (11) can be obtained. On substituting (9), (10) and (11) in (8) one obtains the contribution due to case 1b as

$$\frac{1}{E(X)} \left[1 - \sum_{m=1}^{\hat{I}} P_y(m)\right] \left[1 - \sum_{k=1}^{\hat{I}} P_x(k)\right] \qquad (12)$$

Hence $P_z\ (\hat{I}/Y)$ can be given as the sum of (7) and (12)

$$P_z(\hat{I}/Y) = P_y(\hat{I})\ \frac{1}{E(X)} \sum_{k=\hat{I}}^{x_{max}} (k - \hat{I} + 1)\ P_x(k)$$

$$+ \frac{1}{E(X)} \left[1 - \sum_{m=1}^{\hat{I}} P_y(m)\right] \left[1 - \sum_{k=1}^{\hat{I}} P_x(k)\right] \qquad (13)$$

The probability of an arbitrary pulse on the Z train being caused by a pulse in the Y train can be given by

$$P(^{Y}\text{INTERCEPT}) = \frac{E(X)}{E(X) + E(Y)} \qquad (14)$$

Therefore, if the weighted contribution to $\hat{I}$ due to intervals which begin on pulses contributed by the Y train is defined as $P_z\ (\hat{I})/Y$ then

$$P_z(\hat{I})/Y = \frac{E(X)}{E(X) + E(Y)} \; P_z(\hat{I}/Y)$$

$$P_z(\hat{I})/Y = \frac{E(X)}{E(X) + E(Y)} \left\{ P_y(\hat{I}) \; \frac{1}{E(X)} \sum_{k = \hat{I}}^{x_{max}} (k - \hat{I} + 1) \; P_x(k) \right.$$

$$\left. + \frac{1}{E(X)} \left[1 - \sum_{m = 1}^{\hat{I}} P_y(m) \right] \left[1 - \sum_{k = 1}^{\hat{I}} P_x(k) \right] \right\} \tag{15}$$

Part 2

In Part 2 it is desired to find the probability that an interval of length $(\hat{I})$ occurred in the Z train given that the start of the interval was due to a pulse from the X train. If the simultaneous occurrence of A pulses from the X train and B pulses from the Y train are considered as being contributed by the X train only, then Part 2 becomes a mirror image of Part 1. The result is

$$P_z(\hat{I})/X = \frac{E(Y)}{E(X) + E(Y)} \left\{ P_x(\hat{I}) \; \frac{1}{E(Y)} \sum_{m = \hat{I}}^{y_{max}} (m - \hat{I} + 1) \; P_y(m) \right.$$

$$\left. + \frac{1}{E(Y)} \left[1 - \sum_{k = 1}^{\hat{I}} P_x(k) \right] \left[1 - \sum_{m = 1}^{\hat{I}} P_y(m) \right] \right\} \tag{16}$$

In order to find the total probability, $P_z(\hat{I})$, of a pulse of length $(\hat{I})$ occurring in the Z train, one can sum the contributions due to Part 1 and Part 2 of this derivation. Such a summation, however, will include contributions from both the X train and the Y train for the event when intervals of length $(\hat{I})$ from the X train and intervals of length $(\hat{I})$ from the Y train coincide. This is undesirable and can be corrected by decreasing the contribution of case a, for either Part 1 or Part 2 by an amount which is proportional to the probability of these simultaneous interval occurrences. This can be implemented by changing P(X INTERCEPT) for that portion of the $P_z(\hat{I})/X$ equation corresponding to case 1a to

$$P(X_{INTERCEPT}) = \frac{E(Y)}{E(X) + E(Y)} - \frac{P_x(\hat{I}) P_y(\hat{I})}{E(X) \;\; E(Y)} \tag{17}$$

The first term on the right of equation (17) is the normal probability of X intercept and the second term is the probability of simultaneous occurrences of intervals of length $(\hat{I})$ in the X and Y trains.

The weighted contributions to $P_z(\hat{I})$ due to intervals which begin on the pulses contributed by the X train is therefore

$$P_z(I)/X = \left\{ \frac{E(Y)}{E(X) + E(Y)} - \frac{P_x(\hat{I}) P_y(\hat{I})}{E(X) \;\; E(Y)} \right\}$$

$$\left\{ P_x(\hat{I}) \; \frac{1}{E(Y)} \sum_{m = \hat{I}}^{y_{max}} (m - \hat{I} + 1) \;\; P_y(m) \right\}$$

$$+ \left\{ \frac{E(Y)}{E(X) + E(Y)} \right\} \left\{ \frac{1}{E(Y)} \left[1 - \sum_{k = 1}^{\hat{I}} P_x(k) \right] \left[1 - \sum_{m = 1}^{\hat{I}} P_y(m) \right] \right\} \tag{18}$$

The final equation for $P_z(\hat{I})$ can be obtained by adding equations 15 and 18. In order to find the complete interval probability density $P_z(n)$, equation $P_z(\hat{I})$ would have to be solved for each possible value of n. It is quite easy to see that the maximum range on n (i.e. z_{max}) will be equal to whichever is the smallest of x_{max} or y_{max}.

III CONCLUSIONS

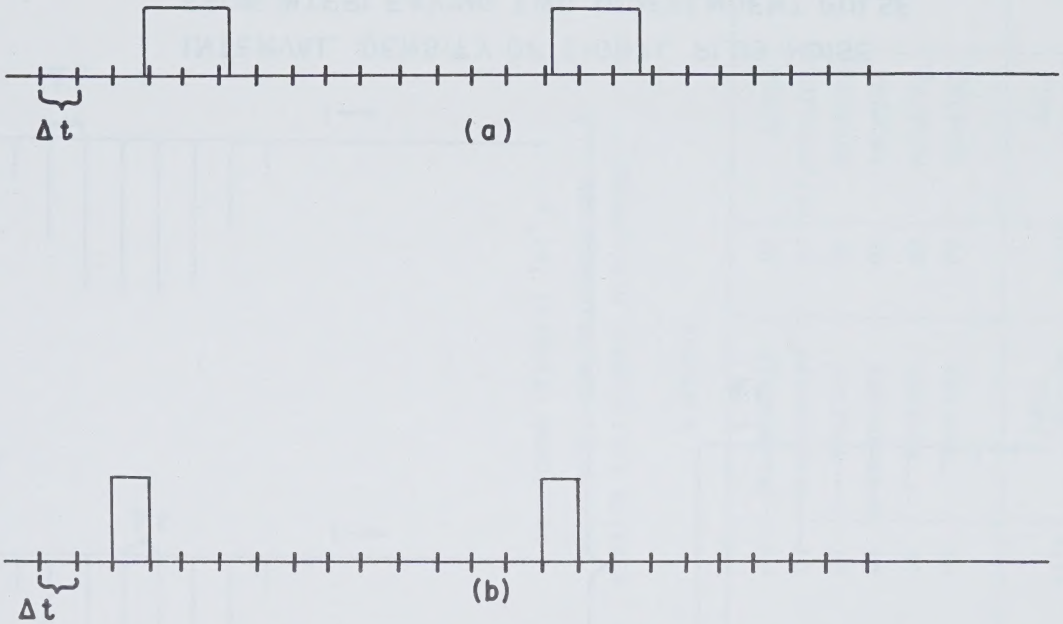
Although the equation for $P_z(\hat{I})$ may seem somewhat formidable at first glance, it is really quite easy to program on a digital computer and requires surprisingly little running time.

It should be noted that the derivation is completely general and is not limited to specific interval densities. The author has checked the equations, by computer simulation, for cases where the X and Y pulse trains were arbitrarily assigned Gaussian, uniform, exponential or point interval probability densities. Runs were made for several different expected values for each pulse train. The regression coefficient between the theoretical and measured $P_z(\hat{I})$ was calculated. Some of these results are tabulated in figure 5. Literally dozens of runs of this type show that no significant bias exists in the results.

The author has successfully utilized these equations in the area of communications theory, however, due to the general nature of the derivation, they would appear to have strong applicability in the field of bio-medical research and queueing theory.

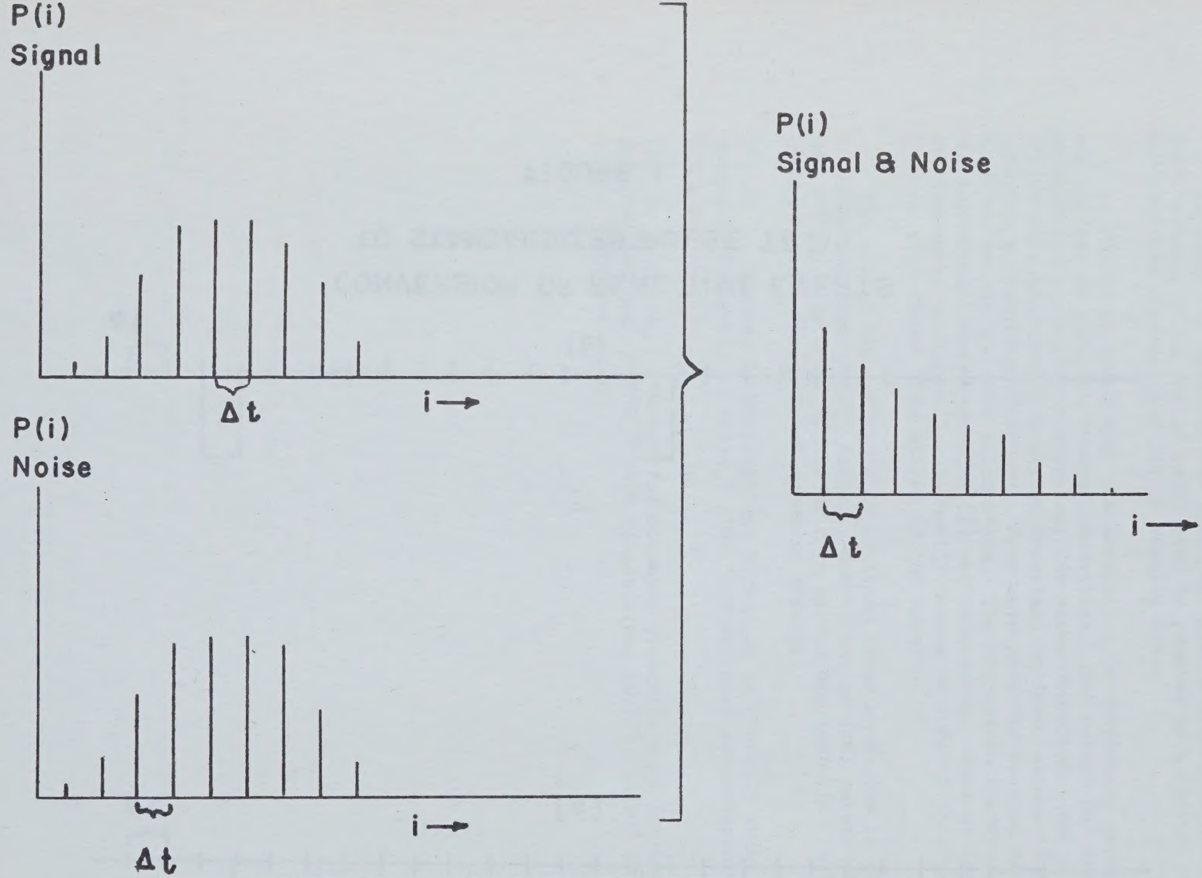
(1) J. H. O'Hara, "A Mathematical Technique to Classify Pulse Position Modulated Telemetry in Noise," 1968 IEEE National Telemetering Conference Rec., p.p. 65-70.

(2) F. F. Hiltz, "A Method for Computer Recognition of Intracellulary Recorded Neuronal Events," IEEE Transactions on Bio-Medical Engineering, April, 1965, p.p. 63-72.



CONVERSION OF REAL TIME EVENTS
TO STANDARDIZED PULSE TRAIN.

FIGURE I.



INTERVAL DENSITY OF SIGNAL PLUS NOISE FROM INTERLEAVING TWO INDEPENDENT PULSE TRAINS.

FIGURE 2

RUN NUMBER	X TRAIN		Y TRAIN		REGRESSION COEFFICIENT
	DENSITY TYPE	E(X)	DENSITY TYPE	E(Y)	
1	UNIFORM	10	UNIFORM	10	1.0137341
2	UNIFORM	10	UNIFORM	20	1.0125339
3	UNIFORM	10	EXPONENTIAL	10	1.0037084
4	UNIFORM	30	UNIFORM	10	0.97324572
5	EXPONENTIAL	10	EXPONENTIAL	10	1.0088680
6	POINT	30	EXPONENTIAL	10	1.0202905

FIGURE 5

REGRESSION COEFFICIENT BETWEEN
EXPERIMENTAL AND CALCULATED
$P_z(\eta)$ INTERVAL DENSITIES

A MODEL FOR PREDICTING PERFORMANCE OF LOW DATA RATE PHASE-LOCKED-LOOP SYSTEMS

T. L. Stewart
Dept. of Electrical Engineering
Bradley University
Peoria, Illinois

K. R. Rao
Dept. of Electrical Engineering
University of Texas
Arlington, Texas

SUMMARY

A nonlinear feedback system called a phase-locked-loop is frequently used to provide the phase information required for coherent detection of digital signals. In the usual case the estimate of the unknown carrier phase is assumed constant over the duration of the signal interval. If the signal duration is much smaller than the loop time constants, this assumption is reasonable. However, at lower data rates it is probable that the carrier phase varies significantly over the detection interval.

In this paper a model is presented which permits the prediction of system performance when the phase estimate varies slowly over the signal interval. This represents a situation between the rapidly varying phase error case recently considered by Hayes and Lindsey(1) and the usual case where the phase error is assumed constant. A hybrid loop(2) model is considered and the probability density function for the phase error and phase error rate are determined by Fokker-Planck techniques. The hybrid loop is considered enabling the performance of a conventional phase-locked-loop or Costas-loop to be readily determined. Also, this approach permits the nonlinear character of the loop to be retained. As the data rate increases, the results reduce to the case assuming a constant phase error.

THE HYBRID LOOP

The block diagram for the hybrid loop is shown in Fig. 1.

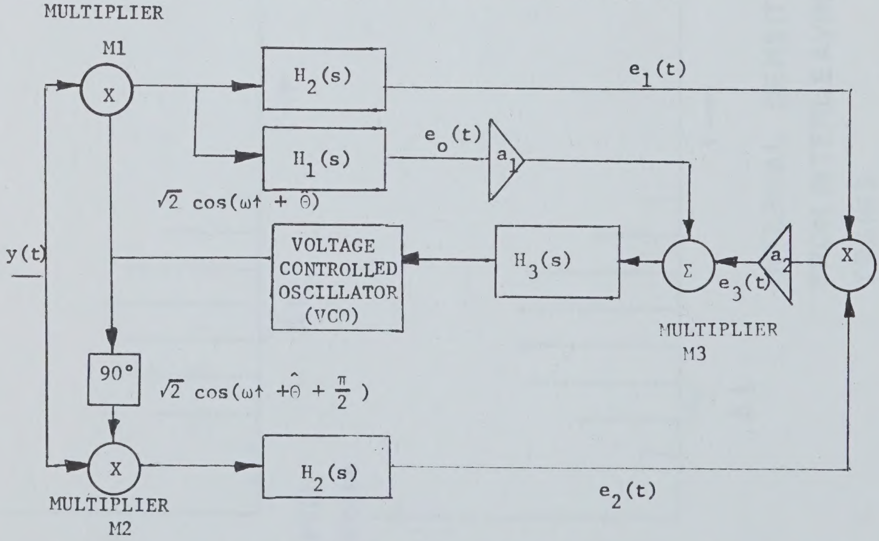

MULTIPLIER
M1
X
$H_2(s)$
$e_1(t)$
$H_1(s)$
$e_o(t)$
a_1
$\sqrt{2}\cos(\omega t + \hat{\theta})$
y(t)
VOLTAGE CONTROLLED OSCILLATOR (VCO)
$H_3(s)$
Σ
$e_3(t)$
a_2
X
MULTIPLIER
M3
90°
$\sqrt{2}\cos(\omega t + \hat{\theta} + \frac{\pi}{2})$
X
$H_2(s)$
$e_2(t)$
MULTIPLIER
M2

Fig. 1 Hybrid loop block diagram

The hybrid loop is a combination of the Costas and conventional phase-lock-loops. By setting the gain a_2 to zero, the hybrid loop reduces to a conventional phase-locked-loop which extracts the phase information from the carrier component of the transmitted signal. Setting a_1 to zero gives a Costas loop mechanization used to obtain a phase estimate directly from the data component of the signal. The hybrid loop combines the error signals from both loop branches to drive the voltage-controlled-oscillator (VCO). The loop filters with transfer functions $H_1(s)$ and $H_2(s)$ remove certain double frequency and modulation components resulting from the multiplier operations and are otherwise assumed flat over the frequency range of the loop. The loop filter with transfer function $H_3(s)$ is the dominate filter with respect to the dynamic characteristics of the system.

The received signal is assumed to be of the form

$$y(t) = s(t) + n(t) \tag{1}$$

where

$$s(t) = \sqrt{2}\, A \sin\left(\omega t + x(t)\left(\cos^{-1} m\right) + \Theta\right) \tag{2}$$

$$n(t) = n_c(t) \cos \omega t + n_s(t) \sin \omega t \tag{3}$$

with $x(t)$, a bipolar-modulated square-wave subcarrier of unit magnitude, $x(t) = \pm 1$; m, the modulation index which adjusts the power allocation between the modulated and carrier portions of $s(t)$; Θ, the carrier phase assumed unknown at the receiver; and $n(t)$, the additive noise component.

The signal $s(t)$ can be expanded into two components representing the carrier and data components of the signal, that is

$$s(t) = \sqrt{2}\, A\, m \sin(\omega t + \Theta) + \sqrt{2}\, A\, (1-m^2)^{1/2}\; x(t) \cos(\omega t + \Theta) \tag{4}$$

In the configuration depicted in Figure 1, the outputs of multipliers M_1 and M_2 after filtering by $H_1(s)$ and $H_2(s)$ are given by

$$e_o(t) = A\, m \sin\phi + \frac{n_c(t)}{\sqrt{2}} \cos\hat{\Theta} - \frac{n_s(t)}{\sqrt{2}} \sin\hat{\Theta} \tag{5}$$

$$e_1(t) = A(1-m^2)^{1/2}\; x(t) \cos\phi + \frac{n_c(t)}{\sqrt{2}} \cos\hat{\Theta} - \frac{n_s(t)}{\sqrt{2}} \sin\hat{\Theta} \tag{6}$$

$$e_2(t) = A(1-m^2)^{1/2}\; x(t) \sin\phi - \frac{n_c(t)}{\sqrt{2}} \sin\hat{\Theta} - \frac{n_s(t)}{\sqrt{2}} \cos\hat{\Theta} \tag{7}$$

Where the phase error ϕ is defined to be $\phi = \Theta - \hat{\Theta}$. The filters $H_1(s)$ and $H_2(s)$ remove all double frequency terms which result from the multiplication. It is further assumed that the bandwidth of $H_1(s)$ is sufficiently narrow to remove the modulated subcarrier $x(t)$ and $H_2(s)$ removes the low frequency component resulting from the carrier. According to this assumption it would be required that the data subcarrier $x(t)$ have insignificant spectral components near the origin.

The output of multiplier M_3 is given by

$$e_3(t) = \frac{A^2 (1-m^2)}{2} \sin 2\phi + \frac{A (1-m^2)^{1/2} x(t) n_c(t) \sin (\phi-\hat{\Theta})}{\sqrt{2}}$$

$$\frac{-A (1-m^2)^{1/2} x(t) n_s(t) \cos (\phi-\Theta)}{\sqrt{2}} - \frac{n_c(t) n_s(t)}{2} \cos 2\hat{\Theta}$$

$$+\left(\frac{n_s(t)^2}{4} - \frac{n_c(t)^2}{4}\right) \sin 2\hat{\Theta} \qquad (8)$$

The error voltage at the input to loop filter $H_3(s)$ is

$$e(t) = a_1 e_o(t) + a_2 e_3(t) \qquad (9)$$

Where the gain factors a_1 and a_2 combine the error signals from the two sections of the loop. The loop differential equation is given by

$$\frac{d\phi}{dt} = -K \; H(p) \; e(t) \qquad (10)$$

Where K is the VCO gain constant and H(p) is the operational transfer function for H_3 assumed to be

$$H(p) = \frac{1+\tau_2 p}{1+\tau_1 p} \qquad (11)$$

corresponding to a second-order loop filter. The loop differential equation becomes

$$\frac{d^2\phi}{dt^2} + \frac{1}{\tau_1} \; [1 + K\tau_2 \; g'(\phi)] \frac{d\phi}{dt} + \frac{K}{\tau_1} g(\phi) = N(t) \qquad (12)$$

Where e(t) has been separated into the two components

$$e(t) = g\;(\phi) + N(t) \qquad (13)$$

with

$$g(\phi) = a_1 Am \sin \phi + \frac{a_2 A^2 (1-m^2)}{2} \; \sin 2\,\phi \qquad (14)$$

$$N(t) = \frac{-K}{\tau_1} [1 + \tau_2 p] \; [a_1 n_o(t) + a_2 n_3(t)] \qquad (15)$$

Where $n_o(t)$ and $n_3(t)$ are the noise components of Eq. 5 and Eq. 8. If we define the state variables

$$y_1(t) = \phi(t)$$

$$y_2(t) = \dot{\phi}(t) \qquad (16)$$

Then the state equations are given by

$$\frac{dy_1}{dt} = y_2$$

$$\frac{dy_2}{dt} = -\frac{1}{\tau_1} [1 + K\tau_2 g'(y_1)]y_2 - \frac{K}{\tau_1} g(y_1) + N(t) \qquad (17)$$

If the bandwidth of N(t) is much wider than the loop bandwidth it is possible to replace N(t) by an equivalent white noise process with correlation function

$$R_{NN}(\tau) \simeq \int_{-\infty}^{\infty} R_{NN}(\tau)\, d\tau \cdot \delta(\tau) \qquad (18)$$

Details of this approximation are given in Stratonovich.(3) With this assumption the process defined by Eq. 17 has a probability density function on $\underline{Y} = (y_1, y_2)$ which satisfies the two-dimensional Fokker-Planck equation

$$\sum_{i=1}^{2} \frac{\partial}{\partial y_i} [K_i (\underline{Y},t) - \frac{1}{2} \sum_{j=1}^{2} \frac{\partial}{\partial y_j} K_{ij} (\underline{Y},t)]\, p(\underline{Y},t) = \frac{\partial p(Y,t)}{\partial t} \qquad (19)$$

where

$$K_i = \lim_{\Delta t \to o} \frac{E \; \Delta y_i | \underline{Y}}{\Delta t} \qquad (20)$$

$$K_{ij} (\underline{Y},t) = \lim_{\Delta t \to o} \frac{E \; \Delta y_i \Delta y_j | \underline{Y}}{\Delta t} \qquad (21)$$

Substituting from Eq. 17 and solving for the coefficients gives

$$K_1 = y_2 \qquad K_2 = -\frac{1}{\tau_1} [1 + K\tau_2\, g'(y_1)]y_2 - \frac{K}{\tau_1}\, g(y_1)$$

$$K_{11} = K_{12} = K_{21} = 0$$

$$K_{22} = \int_{-\infty}^{\infty} R_{NN} (\tau)\, d\tau = N \qquad (22)$$

Substituting these coefficients into Eq. 19 gives

$$\frac{\partial}{\partial y_1} y_2\, p(\underline{Y}) + \frac{\partial}{\partial y_2} [-\frac{1}{\tau_1} (1+K\tau_2\, g'(y_1)]y_2 - \frac{K}{\tau_1}\, g(y_1)]\, p(\underline{Y})$$

$$- \frac{1}{2} \frac{\partial^2}{\partial y_2^2} N\, p(\underline{Y}) = 0 \qquad (23)$$

Although an exact solution to Eq. 23 is unknown, an assumed solution of the form

$$p(\underline{Y}) = C \exp \Big[-\frac{1}{\tau_1 N} y_2^2 - \frac{K\tau_2}{\tau_1 N} g'(y_1)\, y_2^2 - \frac{2K}{\tau_1^2 N} g^{(1)}(y_1) - \frac{K^2\tau_2}{\tau_1^2 N} g(y_1)^2 \Big] \tag{24}$$

where $g'(y_1)$ and $g^{(1)}(y_1)$ represent the derivative and integral of $g(y_1)$ with respect to y_1. Substitution of Eq. 24 into Eq. 23 leaves a remainder term equal to

$$r = -\frac{K\tau_2}{\tau_1 N} y_2^3\, g''(y_1) \tag{25}$$

As the signal amplitude approaches zero, $g''(y_1)$ approaches zero; consequently, it is anticipated that the assumed solution approaches the true probability density function at low power levels, which would be the case for low data-rate systems. Also, the solution is exact for $\tau_2 = 0$ corresponding to a low-pass loop filter.

Substituting for $g(\cdot)$ and N in Eq. 24 gives

$$p(\phi,\dot{\phi}) = C \exp - \alpha_1\dot{\phi}^2 - \alpha_2\dot{\phi}^2(\cos\phi + \rho\gamma\cos 2\phi)$$

$$\cdot \exp \alpha_3(\cos\phi + \frac{\rho\gamma}{4}\cos 2\phi) - \alpha_4(\sin\phi + \frac{\rho\gamma}{2}\sin 2\phi)^2 \tag{26}$$

where

$$\alpha_1 = \frac{\tau_1}{K^2 a_1^2 E} \qquad \alpha_2 = \frac{2\tau_1\tau_2 Am}{Ka_1 E}$$

$$\alpha_3 = \frac{2Am}{Ka_1 E} \qquad \alpha_4 = \frac{\tau_2 A^2 m^2}{E}$$

$$E = \frac{N_o}{2}\left[1 + \rho^2\left(1 - m^2 + \frac{N_o\omega_i}{A^2}\right)\right]$$

$$\rho = \frac{a_1 A}{a_2} \qquad \gamma = \frac{1-m^2}{m} \tag{27}$$

and C is determined from the requirement

$$\int_{-\infty}^{\infty}\int_{-\pi}^{\pi} p(\phi,\dot{\phi})\, d\phi\, d\dot{\phi} = 1 \tag{28}$$

Setting $a_1 = 0$, $m = 0$, $\tau_2 = 0$, Eq. 26 reduces to

$$p(\phi,\dot{\phi}) = C \exp -\beta_1 \dot{\phi}^2 + \beta_2 \cos 2\phi$$

$$\beta_1 = \frac{2\tau_1}{K^2 N_o A^2 \; [1-m^2) + \frac{N_o \omega_i}{A^2} \;]}$$

$$\beta_2 = \frac{A^2}{KN_o \; [N_o \omega_i + A^2]} \tag{29}$$

which corresponds to the Costas loop probability density function found by Didday and Lindsey.(4)

Setting $a_2 = 0$, $\tau_2 = 0$, $a_1 = 1$, gives

$$p(\phi,\dot{\phi}) = C \exp - \frac{2\tau_1^2 \dot{\phi}^2}{K^2 N_o} + \frac{4AM}{KN_o} \cos\phi \tag{30}$$

which is the probability density function for a conventional phase-locked loop with a low-pass loop filter.(5) Finally, setting $\tau_1 = 0$, $\tau_2 = 0$ gives the result

$$p(\phi) = C \exp \frac{4Am}{Ka_1 N_o \; E} \; [\cos\phi + \frac{\rho\gamma}{4} \cos 2\phi]$$

$$E = 1 + \rho^2 (1-m^2 + \frac{N_o \omega_i}{A^2}) \tag{31}$$

which is the probability density function for the hybrid loop found by Stewart.(6)

CALCULATION OF DETECTION PERFORMANCE

The receiver computes

$$d = \int_o^T y(t) \cdot \sqrt{2}\, x(t) \cos (\omega t + \hat{\Theta}(t)) \, dt \tag{32}$$

and decides +1 was transmitted if $d > 0$ and -1 if $d < 0$. The phase estimate $\hat{\Theta}(t)$ is assumed to vary slowly over the duration of the signal interval T. Consequently, it seems reasonable to approximate the resultant phase error by a linear function of the form

$$\phi(t) = \dot{\phi}(t - \frac{T}{2}) + \phi \tag{33}$$

Where ϕ and $\dot{\phi}$ are assumed to be constant over the duration of the signal with the joint probability density given by Eq. 26. Conditioned on ϕ and $\dot{\phi}$ it is easy to show that the resultant error probability is given by

$$P(E|\phi,\dot{\phi}) = \int_{D}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \tag{34}$$

where

$$D = \sqrt{\frac{A^2(1-m^2)}{N_o R}} \qquad \frac{\sin \frac{\dot{\phi}}{2R}}{\frac{\dot{\phi}}{2R}} \cdot \cos\phi \tag{35}$$

and the average error probability can be found from

$$P(E) = \int_{-\infty}^{\infty}\int_{-\pi}^{\pi} P(E|\phi,\dot{\phi})\ p(\phi,\dot{\phi})\ d\phi d\dot{\phi} \tag{36}$$

For the case m = 0 where all the transmitted power is put into the data signal, the Costas-loop is used to provde the phase estimate. Since this system has a phase ambiguity of 180°, some scheme must be used to resolve this difficulty. One such method employs a differentially-encoded data format and the resultant error probability conditioned on ϕ and $\dot{\phi}$ is given by

$$P(E|\phi,\dot{\phi}) = 2P(E|\phi,\dot{\phi})\ [1-P(E|\phi,\dot{\phi})] \tag{37}$$

where $p(E|\phi,\dot{\phi})$ is given by Eq. 34 with m set to zero. The error probability is given by

$$P(E) = \int_{-\infty}^{\infty}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2\ P(E|\phi,\dot{\phi})\ [1-P(E|\phi,\dot{\phi})]\ p(\phi,\dot{\phi})\ d\phi d\dot{\phi} \tag{38}$$

Note that for high data rates, $R/_{\omega_L}$ large,

$$p(\phi,\dot{\phi}) \rightarrow p(\phi)\delta(\dot{\phi}) \tag{39}$$

and Eq. 36 and Eq. 38 reduce to the case where ϕ is assumed to be constant over the duration of the signal.

BIBLIOGRAPHY

1. Hayes, J. F. and Lindsey, W. C., "Power Allocation - Rapidly Varying Phase Error," IEEE Transactions on Communication Technology, Vol. COM-17, April 1969.

2. Stewart, T. L., "A Model Distribution for a Hybrid Phase-Locked-Loop," Proceedings of the Third Asilomar Conference on Circuits and Systems, December 1969.

3. Stratonovich, R. L. Topics in the Theory of Random Noise, Vol. I, Gordon & Breach, New York, 1963.

4. Didday, R. L. and Lindsey, W. C., "Subcarrier Tracking Methods and Communication System Design," IEEE Transactions on Communication Technology, Vol. COM-14, February 1966.

5. Charles, F. J. and Lindsey, W. C., "Some Analytical and Experimental Phase-Locked-Loop Results for Low Signal-to-Noise Ratios," Proc. IEEE, Vol. 54, September 1966.

6. Stewart, T. L., "Analysis of a Hybrid Phase-Locked-Loop," NASA Technical Note, NASA TN D-5666, June 1970.

METHODS OF REDUCING RADAR MULTIPATH TRACKING ERRORS

PEYTON Z. PEEBLES, JR.
University of Tennessee
Knoxville, Tennessee 37916

ABSTRACT

Three different methods, with four variations on one of the methods, have been investigated for reducing radar angle tracking errors due to multipath. The six possibilities are reduced to three by consideration of implementation complexity. Two of the three methods simulated on a digital computer have shown that large reductions in angle errors are possible as compared to a normal monopulse system.

INTRODUCTION

The purpose of this paper is to briefly describe three methods for reducing radar angle tracking errors due to multipath signals. The intent here is primarily to present results, using the three methods, of digital computer simulations of a radar with the usual noise and multipath present. However, references to the literature are also given so that the reader may obtain detailed discussions of the various methods.

The first method shall be called the complex angle (CA) method. It is basically a special method of signal processing the received signals; it assumes some knowledge (both electrical and geometrical) exists about the reflecting surface around the radar. The method was conceived by Sherman [1] and can be classed as a special form of single-target signal processor even though the method was an outgrowth of studies of a two-target problem. Detailed discussions are given in [2].

The remaining two methods treat the multipath problem as a two-target problem. One shall be called the nonlinear method. It was suggested by Peebles and Berkowitz [3] and details of the signal processor are given in [3]. Another method which, in theory, allows for (N -1) multipath components has been developed and employs a linear-adaptive signal processor. It will here be called the adaptive method. Detailed discussions of this method are given in [4].

We shall consider only amplitude comparison monopulse, that is, beam patterns of the radar are real functions. Furthermore, discussion is limited to the elevation angle tracking problem, since most multipath problems occur in this tracking axis. Finally, for simplicity, a flat-surface geometry is assumed as illustrated in Fig. 1, which also defines notation.

THE CA METHOD

Consider first the case of no noise. If multipath is not present the elevation angle E_1, relative to the boresight axis, is given by a real quantity

$$E_1 = \frac{1}{p}\frac{D}{S}, \tag{1}$$

where p is a proportionality constant, D is the difference (error) channel voltage and S is the sum (reference) channel voltage. When single reflection multipath is present both D and S have two signal components so that the right side of (1) is complex. The left side becomes a complex angle estimate, the real part being the usual angle estimate formed by a monopulse system with multipath.

Analysis with multipath and no noise easily shows that (Sherman [1])

$$\frac{1}{p}\frac{D}{S} = -E_o + E\,\frac{1 - g\Gamma e^{j\Delta\phi}}{1 + g\Gamma e^{j\Delta\phi}}, \tag{2}$$

where

$$g = \frac{G_s\,(-E-E_o)}{G_s\,(E-E_o)} \tag{3}$$

$$\Delta\phi = \frac{-4\pi h}{\lambda}\ \sin\,(E), \tag{4}$$

h is the radar elevation, λ is wavelength, Γ is the complex reflection coefficient of the surface as defined in Kerr [5], and $G_s(\cdot)$ is the one-way voltage pattern of the sum channel.

PROCESSOR

The first key to the CA method is to implement a receiver to measure both in-phase and quadrature voltages resulting from the ratio of the left side of (2). Thus the system has a measurement of the complex angle corresponding to the left side of (2) with noise. The second key to the CA method is to treat the right side of (2) as a computed quantity. The function can be computed by iterating values of E until a value $\hat{E}$ is found that causes the equality. Without noise, the value of $\hat{E}$ will equal the true angle E.

If noise is present, the iteration selects the $\hat{E}$ that minimizes the magnitude of the difference between the left side (measured value of complex angle) of (2) with noise and the right side (computed) value of (2) without noise. Of course, such a signal processor must know the system beam pattern $G_s(\cdot)$ as well as other parameters such as Γ, h, E_o and λ. The effects of not knowing exactly some of these parameters may be evaluated [2]. However, we here assume that the radar knows all required parameters exactly.

SIMULATION

The CA method was simulated on a digital computer for an assumed system having a sum pattern beamwidth of approximately 1.0 degree, ratio h/λ of 100, and Γ determined by sea water as described in Kerr [5]. The problem is essentially that used by Sherman [1]. The boresight axis was elevated one half beamwidth off the horizon and off-axis elevation angle estimates were obtained for target elevation angles from 18 mr. down to 0.5 mr. Data are shown in Fig. 2 for the bias of angle error using two signal-to-noise (S/N) ratios. Also shown is the bias error that would be obtained with a normal system without the CA method. Bias error is reduced by up to 100 times by the new technique. Bias errors fluctuate in sign and decrease as (S/N) ratio increases leading to the conclusion that the CA method is actually unbiased.

Fig. 3 shows a curve of the standard deviation σ_E of angle error for a number of (S/N) ratios. The nearly fixed trend shows that σ_E is proportional to $(S/N)^{-1/2}$. Also shown is σ_E for a normal system. Improvements of around 3 to 1 are achieved over most of the angles with the CA method.

THE NONLINEAR METHOD

Peebles and Berkowitz [3] have described a two-target nonlinear signal processor. By considering multipath as a second target such a processor may be used to make off-axis elevation angle estimates by fixing the boresight axis a half beamwidth above the horizon as in the CA method.

Using the same system parameters as in the CA method above, the nonlinear method was simulated on a digital computer. Results are summarized* in Fig. 4. It is seen that large angle errors occur at periodic target elevation angles. Since the results are generally poor, especially in comparison to the other methods discussed here, the nonlinear method is not considered useful and is not considered further.

THE ADAPTIVE METHOD

It can be shown that, for either an N-target problem [3] or a multipath problem with (N - 1) reflection paths [4], the signal e_m emerging from the mth beam of M receive beams is

$$e_m = \sum_{n=1}^{N} A_n G_{mn} \tag{5}$$

where A_n is an equivalent (complex) amplitude of the wave arriving from direction n and G_{mn} is the one-way receive-beam voltage pattern evaluated in the nth path's direction. Thus, multiple-target analyses may be applied to the multipath problem.

The adaptive method is based upon expanding the beam patterns in (5) into Taylor series involving M complex equations in the 4 N real unknowns, which are the two angular directions of each received ray as well as the amplitude and phase of the A_n. The system of equations may then be solved for the unknowns. The approach is similar to that described in [3], however, one difference forms the key to the adaptive method.

If we assume a priori knowledge of the approximate directions of the N signals, we may expand G_{mn} into Taylor series about these N directions. We form, in effect, (N - 1) expansion axes other than the boresight axis which applies to the main ray. Fig. 1 illustrates such an axis, called the lower expansion axis (LEA), for the single-reflection multipath case. Using a K-degree Taylor series expansion in (5) gives

$$e_m = \sum_{n=1}^{N} A_n \sum_{k=0}^{K} \sum_{\ell=0}^{k} \alpha_{mnk\ell}\, x_n^{k-\ell}\, y_n^{\ell} \tag{6}$$

where the $\alpha_{mnk\ell}$ are the Taylor coefficients and x_n and y^n are the angles relative to the expansion axes.

With some manipulation it can be shown [4] by solving (6) that when N = 2 (one multipath reflection) we require K = 1 and M = 6 beams. A similar analysis, assuming all beams are factorable, with each factor expanded into a K degree Taylor series, leads to K = 1 and M = 8 beams required. Both these cases assume transmission and reception of a single polarization. When dual polarization is assumed it is found that again K = 1 for factorable or non-factorable beams. The beams required in the two cases are: M = 6 for non-factorable and M = 8 for factorable.

---

*Fig. 4 was provided by Mr. R. Ottinger of RCA.

Thus we have four variations of the method depending on the assumptions of polarization and factorability. Since the one-polarization, non-factorable case is the least complex and most likely one to occur in practice it was selected for detailed evaluation.

SIGNAL PROCESSOR

We now develop a signal processor for the important case of a single reflection. Define

$$Z_{nk\ell} = A_n x_n^{k-\ell} y_n^{\ell}, \quad n = 1, 2, \tag{7}$$

so that (6) may be written in matrix form as

$$[e] = [\alpha][Z] \tag{8}$$

with

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_6 \end{bmatrix}, \quad [Z] = \begin{bmatrix} Z_1 \\ --- \\ Z_2 \end{bmatrix}, \quad [Z_n] = \begin{bmatrix} Z_{n00} \\ Z_{n10} \\ Z_{n11} \end{bmatrix} \tag{9}$$

for n = 1, 2. Clearly from (8) we have

$$[Z] = [\alpha]^{-1}[e] \tag{10}$$

The matrix $[\alpha]$ is element-ordered to be consistent in (8) and can be argued to have an inverse, since beam patterns may be chosen almost arbitrarily. Because the e's are measureable and $[\alpha]^{-1}$ is known we may consider $[Z]$ known. Finally from (7) we solve for unknowns.

$$x_n = Z_{n10}/Z_{n00} \tag{11}$$

$$y_n = Z_{n11}/Z_{n00} \tag{12}$$

$$A_n = Z_{n00}, \tag{13}$$

all for n = 1, 2.

The adaptive signal processor must first form a set of Z voltages by weighting the beam output e's using the matrix $[\alpha]^{-1}$ from (10). Ratios of the Z's are then taken to find the target angles from (11) and (12).

The above analysis assumes no noise. If the e's are noisy a similar analysis leads to a signal processor which again weights (noisy) e's by $[\alpha]^{-1}$ to form (noisy) Z's according to (10). The angle estimates are then given by the real parts of the ratios of the noisy Z's as in (11) and (12).

The signal processor is adaptive in the sense that the matrix $[\alpha]^{-1}$ must change as a function of target elevation angle. This is easily seen from the fact that the elements $\alpha_{mnk\ell}$ of $[\alpha]$ for n = 1 are fixed (the boresight axis does not change in relation to the directions of the 6 beams because the beam cluster tracks the main ray) while the elements for n = 2 vary with elevation (the LEA has an angle of approximately $-2E_1$ (actually $-2E_0$) relative to the BSA from Fig. 1).

SIMULATION

In order to simulate the adaptive method choices must be made of the beam pattern form and the beam spatial directions. A pattern form which approximates a parabolodial antenna illuminated by a horn offset from the focus is

$$G(x,y) = G_o \frac{2J_1(\xi)}{\xi} \tag{14}$$

where

$$\xi = \frac{3.24}{\theta_B} (x^2 + y^2)^{1/2}, \tag{15}$$

θ_B is the one-way 3dB beamwidth, G_o is the maximum gain, and $J_1(\cdot)$ is the Bessel function of the first kind of order one.

The 6 beams were clustered relative to the BSA according to Fig. 5. The cluster is recognized as the usual monopulse 4-beam arrangement with two new beams (horns) added. θ_B was chosen such that a normal monopulse sum channel formed from the inner 4 beams was approximately one degree. The choice of beam cluster then represents a simple addition of two beams to a normal monopulse system.

Fig. 6 illustrates bias and standard deviation of elevation angle errors from simulation of the adaptive method for a 40dB (S/N) ratio. Noise was assumed added to the e's prior to the weighting matrix $[\alpha]^{-1}$. Each point was obtained by averaging 100 errors using Monte Carlo methods. Although Fig. 6, which is for a simulated tracking case, cannot be compared directly to the normal monopulse curves of Figures 2 and 3, which are for off-axis angle estimation, a crude comparison may be made. This is because it is felt that tracking errors in a normal monopulse case should not be too different from the off-axis data in order of magnitude. Such a comparison shows that large bias improvements result with the adaptive method. Data for other (S/N) ratios [4] show that bias is a decreasing function of increasing (S/N) ratio. These facts, along with sign fluctuations, lead to the conclusion that the method is approximately unbiased.

Standard deviation of error is larger than normal monopulse at nearly all elevation angles. However, it can be shown that the error compares favorably with the minimum possible error for two-target systems of this type [4].

ACKNOWLEDGEMENTS

The writer gratefully acknowledges the assistance of Mr. R. Udicious who programmed the computer for the CA method simulation, Dr. R. Johnson who programmed the simulations of the nonlinear and adaptive methods, Messrs. L. Goldman and R. W. Ottinger who assisted in obtaining simulation results.

REFERENCES

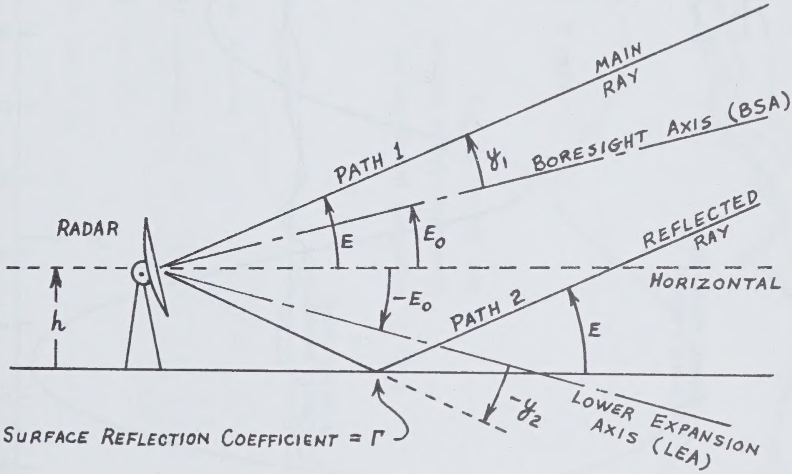
[1] Sherman, S., "The Use of 'Complex Indicated Angles' in Monopulse Radar to Locate Unresolved Targets," Proc. of the Nat. Elect. Conf., vol. 22, 1966, p.243.

[2] Peebles, Jr., P. Z., and Goldman, Jr., L., "Radar Performance with Multipath Using the Complex Angle," To be published, IEEE Trans. on AES, Jan., 1971.

[3] Peebles, Jr., P. Z., and Berkowitz, R. S., "Multiple-Target Monopulse Radar Processing Techniques," IEEE Trans. on AES, Nov., 1968, pp. 845-854.

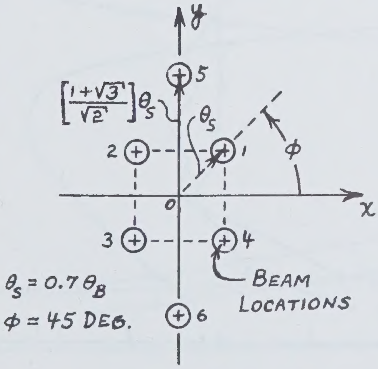
[4] Peebles, Jr., P. Z., "Multipath Angle Error Reduction Using Multiple-Target Methods," to be published, probably in IEEE Trans. on AES during 1971.

[5] Kerr, D. E. (Ed.), Propatation of Short Radio Waves, vol. 13, M. I. T. Rad. Lab. Series, Boston Tech, Publishers, Inc., Lexington, Mass. 1964.



MAIN RAY
BORESIGHT AXIS (BSA)
PATH 1
ψ_1
RADAR
E
E_0
REFLECTED RAY
HORIZONTAL
h
$-E_0$
PATH 2
E
$-\psi_2$
LOWER EXPANSION AXIS (LEA)
SURFACE REFLECTION COEFFICIENT = Γ

Fig. 1 Multipath Geometry



y
5
$\left[\frac{1+\sqrt{3}}{\sqrt{2}}\right]\theta_s$
θ_s
ϕ
2
1
0
x
3
4
BEAM LOCATIONS
$\theta_s = 0.7\,\theta_B$
$\phi = 45$ DEG.
6

Fig. 5 Beam Locations

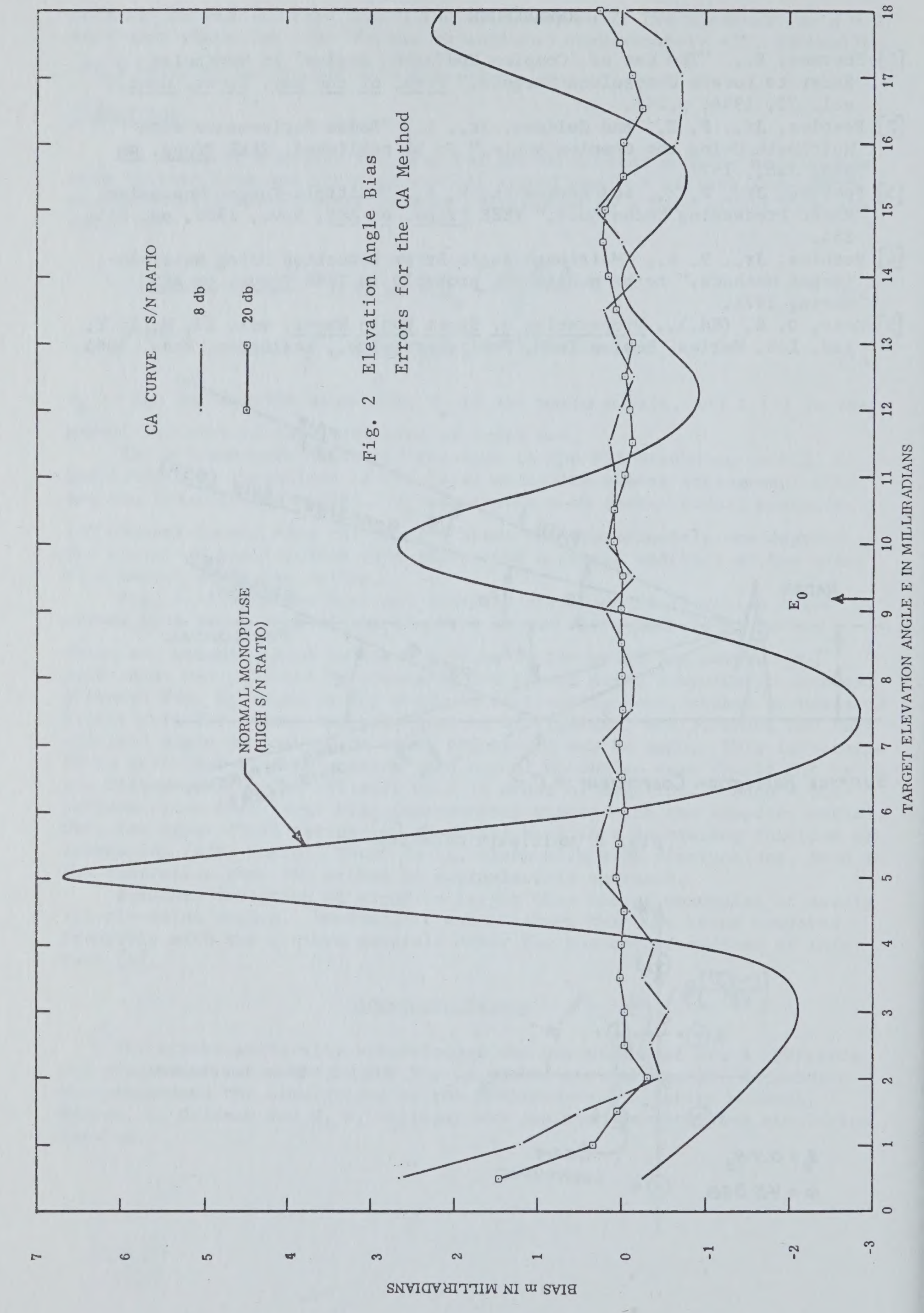


Fig. 2 Elevation Angle Bias Errors for the CA Method

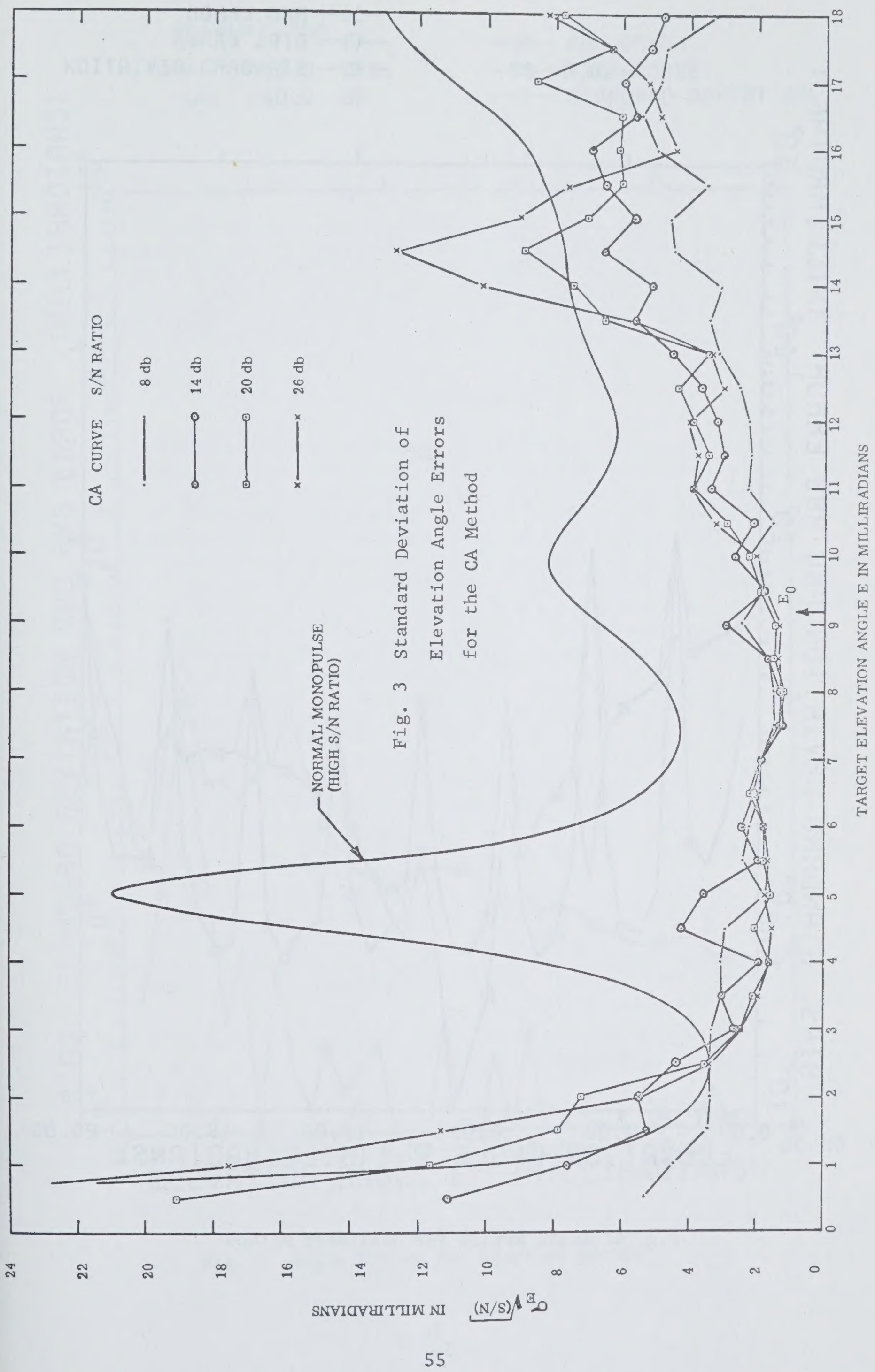


Fig. 3 Standard Deviation of Elevation Angle Errors for the CA Method

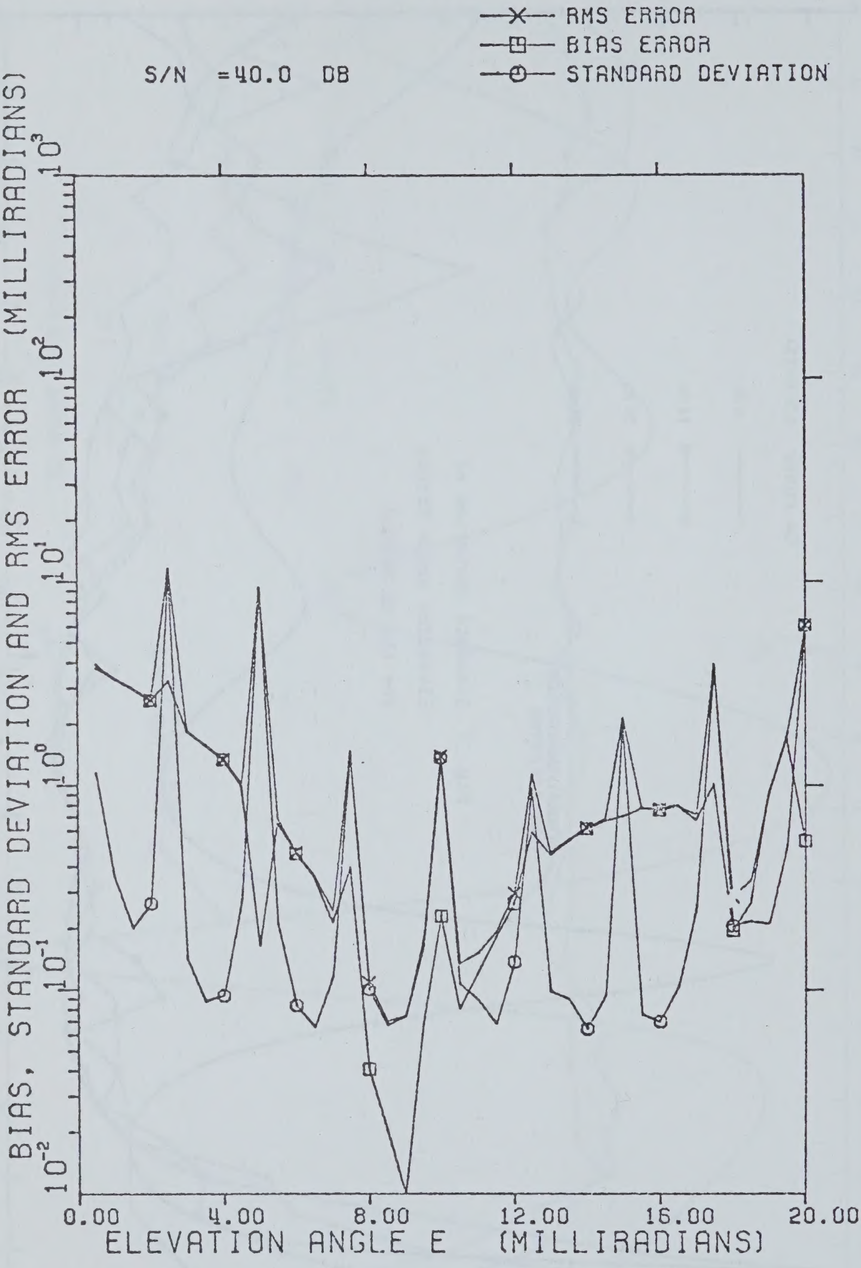


Fig. 4 Angle Errors for Nonlinear Method

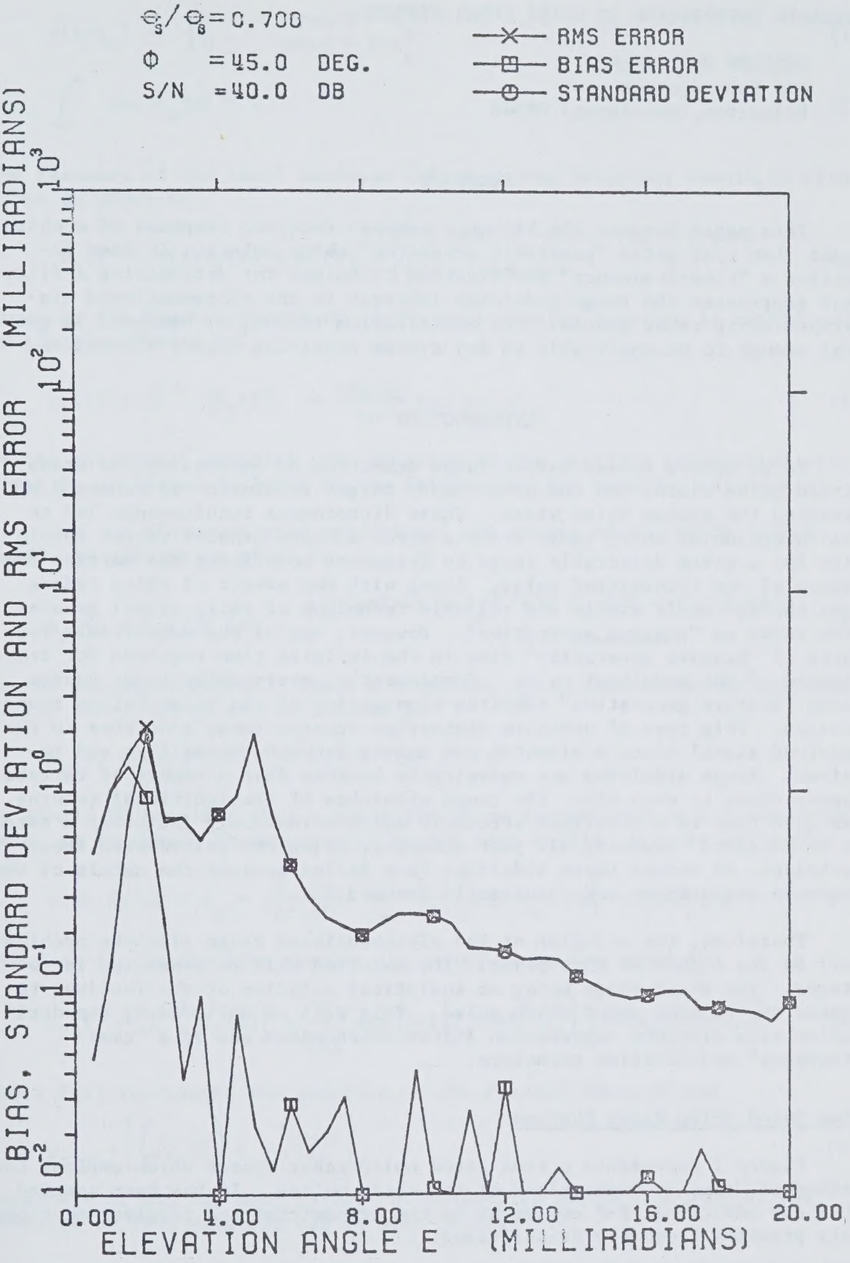


Fig. 6 Angle Errors for Adaptive Method

SIDELOBE SUPPRESSION IN CHIRP RADAR SYSTEMS*

STEPHEN HONICKMAN
RCA Laboratories
Princeton, New Jersey 08540

ABSTRACT

This paper derives the hitherto unknown receiver response of a chirp radar that time gates "passively generated" chirp pulses. It then describes a "time-frequency" modification technique for determining a filter that suppresses the range sidelobes inherent in the aforementioned mismatched chirp radar system. The modification technique, however, is general enough to be applicable to any system requiring signal alteration.

INTRODUCTION

Pulse radars extend target range detection by increasing the transmitted pulse width. On the other hand, target resolution is enhanced by reducing the system pulse width. These dichotomous requirements led to the invention of chirp radar systems which achieve greater target resolution for a given detectable range by frequency modulating the carrier frequency of the transmitted pulse. Along with the advent of chirp radars came the extremely simple and reliable technique of chirp signal generation known as "passive generation". However, one of the undesirable features of "passive generation" lies in the infinite time required for transmission of the resultant pulse. Consequently, every chirp radar system using "passive generation" requires time gating of the pulse before transmission. This type of waveform distortion creates range sidelobes in the received signal since a mismatch now exists between transmitter and receiver. Range sidelobes are undesirable because when a number of targets appear close to each other the range sidelobes of the individual returns can give rise to a resultant structure which erroneously indicates a target to be present. However, all past attempts, using the paired-echo theory technique, to reduce these sidelobes have failed because the nature of the response was unknown and incorrectly assumed.

Therefore, the solution of the aforementioned range sidelobe problem will be the object of this paper. The solution will be presented in two stages. The first stage being an analytical solution of the receiver response to the time gated chirp pulse. This will be followed by the derivation of a sidelobe suppression filter which makes use of a "time-frequency" modification technique.

Time Gated Chirp Radar Systems

Figure 1 represents a time gated chirp radar system which employs the method of "passive generation" of the chirp pulses. It has been assumed that any additional R-F equipment in the transmitter and receiver sections only produces frequency translations.

The input to this system, which is the unit impulse function, can be described by the following equations

---

*This paper has been written in partial fulfillment of the Ph.D. requirements at Newark College of Engineering.

$$\delta(t-t_o) = \begin{cases} \infty \; ; & \text{when } t = t_o \\ 0 \; ; & \text{when } t \neq t_o \end{cases} \tag{1}$$

$$\int_{-\infty}^{\infty} \delta(t-t_o)dt = 1 \; . \tag{2}$$

The response of the ideal bandpass filter, in the frequency domain, to this input is given by

$$E_1(f) = \delta(f)H_b(f) = H_b(f) = \begin{cases} \frac{1}{\Delta} \; ; & \text{when } f_o-\Delta/2 < |f| < f_o+\Delta/2 \\ 0 \; ; & \text{elsewhere} \end{cases} \tag{3}$$

(because the Fourier Transform of the unit impulse function is unity) and in the time domain by,

$$e_1(t) = \mathfrak{F}^{-1}\left[E_2(f)\right] = \frac{\sin\pi\Delta t}{\pi\Delta t} \sin 2\pi f_o t \; . \tag{4}$$

In the subsequent material, the solution to the positive exponential will first be evaluated followed by the solution to the negative exponential.

For the aforementioned conditions, the output of the chirp filter, $E_2(f)$, is seen to be given by

$$E_2(f) = E_1(f)H_c(f) = \begin{cases} \frac{1}{\Delta} \exp(-i\pi\left[f-f_o\right]^2/k); & f_o-\Delta/2 < f < f_o+\Delta/2 \\ 0 \; ; & \text{elsewhere} \end{cases} \tag{5}$$

in the frequency domain. Where $k = \Delta/T$ and $H_c(f)$ represents the spectrum of the chirp filter and is defined as

$$H_c(f) = \exp\left[-i\pi(f-f_o)^2/k\right] \; ; \quad \text{for } f > 0 \tag{6}$$

It should be recalled that a quadratic phase factor implies a linear time delay vs frequency characteristic in the time domain; i.e. by definition,

$$\text{time delay} = t_d = -\frac{1}{2\pi}\frac{\partial\varphi}{\partial f} = -\frac{1}{2\pi}\frac{\partial}{\partial f}\left[-\pi(f-f_o)^2/k\right] = (f-f_o)/k \; . \tag{7}$$

See Fig. 2. In the time domain, the waveform is described by

$$e_2(t) = \mathfrak{F}^{-1}\left[E_2(f)\right] = \frac{1}{\sqrt{2D}} \exp\left(2\pi i\left[f_o t+kt^2/2\right]\right)\left(Z^*(v_2)-Z^*(v_1)\right) \tag{8}$$

where $Z(v)$ represents the solution of the Fresnel Integral and

$$v_{1,2} = \left[\left(\sqrt{2\Delta/T}\right) t \pm \sqrt{D/2}\right] \tag{9}$$

where $D = \Delta T$ and has been termed the Dispersion Factor. A plot of the envelope of this waveform can be seen in Fig. 3.

These results have already been presented in reference 1 and have been reproduced here for convience only. The work which follows represents the contributions of this paper.

It is obvious from Fig. 3 that such a waveform cannot be used by a pulse radar because its infinite time extent creates a zero pulse repetition rate. However, this situation can be remedied by the time gate shown in Fig. 1. The introduction into the system of this time gating process, however, produces a system which no longer possesses matched characteristics. As a result, the de-chirp filter response contains range sidelobes that the paired-echo theory, introduced in reference 1, has been unable to reduce or eliminate because the nature of the de-chirp filter response has been incorrectly assumed to contain no FM components.

Thus, the first step in the solution of the range sidelobe suppression problem will be the analytical derivation of the de-chirp filter response. This will then be followed by the derivation of a sidelobe suppression filter.

De-chirp Filter Response

The time gating network can be described in equation form by

$$r(t) = \begin{cases} 1 ; & |t| < T/2 \\ 0 ; & \text{elsewhere} \end{cases} \tag{10}$$

or

$$R(f) = \frac{T\sin\pi\Delta T}{\pi\Delta T} \tag{11}$$

Therefore, the transmitter output is given by

$$E_3(f) = \left[r(t)e_2(t)\right] = R(f)*E_2(f) = \frac{1}{\Delta}\int_{\alpha_1}^{\alpha_2} \exp -i\pi\left[f-\alpha-f_o\right]^2/k \,\frac{\sin\pi\alpha T}{\pi\alpha}\,d\alpha \tag{12}$$

where $\alpha_2 = f_o - \left[f_o - \Delta/2\right]$ and $\alpha_1 = f_o - \left[f_o + \Delta/2\right]$

Assuming that the received signal differs from the transmitted in amplitude only, the de-chirp output can be written as

$$E_4(f) = \frac{\exp\left[i\pi\beta^2/k\right]}{\Delta}\int_{\beta-\Delta/2}^{\beta+\Delta/2} \exp\left[-i\pi[\beta-\alpha]^2/k\right] \frac{\sin\pi\alpha T}{\pi\alpha}\,d\alpha = E_3(f)H_d(f) \tag{13}$$

where $\beta = (f-f_o)$ and $H_d(f)$ represents the spectrum of de-chirp filter which is given by

$$H_d(f) = \exp\left[i\pi\beta^2/k\right] \qquad \text{for } f > 0 \tag{14}$$

Since the integral in equation (13) does not exist in standard form an ingenious technique was evolved that makes use of the following mathematical properties

$$\frac{\partial}{\partial\beta}\int_{f_1(\beta)}^{f_2(\beta)} F(\alpha,\beta)\,d\alpha = \int_{f_1(\beta)}^{f_2(\beta)} \frac{\partial F(\alpha,\beta)}{\partial\beta}\,d\alpha + F(f_2(\beta),\beta)\frac{\partial f_2(\beta)}{\partial\beta} - F(f_1(\beta),\beta)\,\frac{\partial f_1(\beta)}{\partial\beta} \tag{15}$$

Denoting result in equation (15) by $E_5(\beta,\alpha)$, it follows that

$$\mathcal{F}^{-1}\left[E_5(\beta,\alpha)\right] = e_5(t) = \frac{2\pi t}{i}\mathcal{F}^{-1}\left[E_4(\beta,\alpha)\right] = \frac{2\pi t}{i}\, e_4(t) \qquad (16)$$

The application of equation (15) to equation (13) yields, after rather extensive and complicated manipulations, an expression which has a standard mathematical form. Due to its complexity and length only the results are presented here; however, the derivation in its entirety can be obtained from the author. The results of this analysis yield the following equations,

$$e_4(t) = e_{41}(t) + e_{42}(t) + e_{43}(t) \qquad (17)$$

where

$$e_{41}(t) = \frac{Z^*(\sqrt{2D})}{\sqrt{2}} \exp\left(2\pi i\left[f_o t - kt^2/2 + 1/8\right]\right) \frac{\sin\pi\Delta T}{\pi\Delta t} \qquad (18)$$

and

$$e_{42}(t) = \left[\frac{Z^*(\sqrt{2k}\ t) - Z^*(\sqrt{2k}\ (t-T))}{\sqrt{2k}}\right] \frac{\exp\left(2\pi i\left([f_o - \Delta/2]t + 3/8\right)\right)}{2\pi\sqrt{D}\ t} \qquad (19)$$

and

$$e_{43}(t) = \left[\frac{Z^*(\sqrt{2k}\ (t+T)) - Z^*(\sqrt{2k}\ t)}{\sqrt{2k}}\right] \frac{\exp\left(2\pi i[(f_o + \Delta/2)t - 1/8]\right)}{2\pi\sqrt{D}\ t} \qquad (20)$$

At this juncture, having ascertained the response due to the positive exponential, the next step will be the evaluation of the response due to the negative exponential. However, for the case at hand, the response cannot be found by simply taking the complex conjugate of eq. (17) because of the existence of a quadratic phase factor in the chirp and de-chirp filter d definitions. The quadratic phase factor implies that in order for the phase angle to have odd symmetry the expression for the chirp filter must appear as follows

$$H_c(f) = \exp[i\pi(f+f_o)]^2 \ ; \quad \text{for all negative frequencies.} \qquad (21)$$

This expression, as well as that of the de-chirp filter, displays two changes from its definition in the positive frequency domain. The first change being the standard complex conjugation and the second in the sign of the quadratic term; i.e. $f-f_o$ becomes $f+f_o$. The second change can be incorporated into the results by making use of the following property of Fourier Transforms

$$\mathcal{F}^{-1}[F(f\pm u)] = \exp(\mp 2\pi i u t)F(t) \qquad (22)$$

Therefore, the quantity f_o in all previous equations must be changed to $(-f_o)$. This change takes place in addition to the complex conjugation process.

By applying the aforementioned changes to eq. (17) and, then, combining the result with eq. (17) yields, after some extensive mathematical manipulations, the following expression for the de-chirp filter output waveform.

$$e_o(t) = \sqrt{2}\, A\, \frac{\sin\pi\Delta t}{\pi\Delta t} \cos 2\pi(f_o t - kt^2/2 + a) + \cos 2\pi f_o t \left[\frac{\cos\pi\Delta t}{\pi\Delta t}(S_1 - V_2) + \frac{\sin\pi\Delta t}{\pi\Delta t}(S_2)\right] + \sin 2\pi f_o t \left[\frac{\cos\pi\Delta t}{\pi\Delta t}(S_2 - V_1) - \frac{\sin\pi\Delta t}{\pi\Delta t}(S_1)\right] \tag{23}$$

where $Ae^{-ic} = Z^*(\sqrt{2D})$ and $a = (1/8) - c'$, i.e. $c' = c/2\pi$

$$S_{1,2} = \frac{1}{2}(R_{+}-X) \text{ and } V_{1,2} = (P \pm Q) \tag{24}$$

where

$$P + iQ = Z(t_1) - Z(t_1 - T_1) \text{ and } R + iX = Z(t_1 + T_1) - Z(t_1 - T_1) \tag{25}$$

where $t_1 = \sqrt{2k}\, t$ and $T_1 = \sqrt{2k}\, T$.

A plot of the waveform with $T = 3\mu$sec, $\Delta = 10$ MHz and $f_o = 30$ MHz can be seen in Fig. 4. The figure indicates that the first sidelobe is only -13 dB below the main lobe.

Since this result obviously possesses an F-M structure, it is impossible to obtain advanced and delayed replicas of this waveform which possess the phase structure required by the paired-echo theory for sidelobe suppression. In the past, however, all attempts at sidelobe suppression made use of the paired-echo theory technique because the result in eq. (23) was unknown.

Since the time gating of the chirp pulse has been responsible for the signal structure in eq. (23), it would be reasonable to assume that without the time gate (i.e. $T \to \infty$) eq. (23), in the limit, should yield the result given in ref. 1. Thus, with $T \to \infty$, $k = \Delta/T \to 0$, eq. (23) yields

$$e_4(t) = \frac{\sin\pi\Delta t}{\pi\Delta t} \cos 2\pi f_o t \tag{26}$$

as expected.

Spectrum Of The De-chirp Response

Having ascertained the de-chirp response, the next step is the evaluation of a sidelobe suppression filter. Unfortunately, the complexity of eq. (23) does not permit an analytic solution. Therefore, the remainder of this paper is devoted to a description of the derivation of a sidelobe suppression filter by digital techniques. Upon completion of the aforementioned derivation, a generalized approach is formulated which can be used for numerous problems.

The first step in the analysis requires a knowledge of the frequency spectrum. Since it is to be determined by digital techniques a number of questions must first be answered. First, what is the sampling rate? And, second, given a sampling rate, How does one know that the resulting spectrum is that which is required?

The first question can be answered by examining the time domain waveform. Figure 4 indicates that the envelope appears to have a sinx/x structure. This implies that the spectrum possesses a rectangular waveshape whose bandwidth approximates that of the overall system. Therefore, with a carrier frequency of 30 MHz, it is reasonable to assume that the highest frequency component lies in the vicinity of 35 MHz.

The second question has been answered by such people as Tuckey, Cooley and etc. They contend that one achieves the required spectrum when the spectra obtained from sampling rates differing by a factor of 2 are negligibly different.

With this information, one is now able to write a computer program which makes use of the Fast Fourier Transform (FFT) to evaluate the spectrum of the response. In order to satisfy the sampling theorem and the FFT program, the following sampling rates are to be used,

$$T_{s_1} = 6.25 \text{ ns} \quad T_{s_2} = 3.125 \text{ ns} \quad \text{and} \quad T_{s_3} = 1.5625 \text{ ns} \tag{27}$$

Using these sampling rates in the aforementioned computer program gives rise to spectra which are in the Tukey, Cooley sense, negligibly different from each other. The resulting spectrum for T_{s_3} and the one which is to be used in all subsequent analyses can be seen in Fig. 5. This spectrum clearly indicates the validity of the previous assumptions regarding its structure.

Derivation Of The Sidelobe Suppression Filter

This section is devoted to the derivation of a filter whose function is the suppression of the range sidelobes existant in the de-chirp response. The technique, which is going to be used in the derivation, makes use of simultaneous alterations of the time domain and frequency domain waveshapes to yield the required filter.

The first step in the analysis requires dividing the two waveshapes into desirable and undersirable regions. In this particular case, for the frequency spectrum, this implies dividing the spectrum into the region lying inside the system bandwidth and a region outside and, for the time domain waveform, it means dividing the signal into the region lying inside the main lobe and a region outside.

At this point, a decision must be made as to the first alteration. For this particular case the most appealing choice is the reduction to -40 dB below the main lobe of all sidelobes existant in the time domain signal. This alteration allows determination of a new spectrum via use of the FFT which will indicate those portions of the original spectrum responsible for these sidelobes and at the same time gives an indication of the filter necessary for their reduction. As one would expect, this filter is non-realizable. Therefore, in order to proceed with the analysis it is necessary to concentrate on those charges which are realizable and eliminate all others.

The result of this alteration clearly shows that many of the sidelobes stem from that section which exists within the system bandwidth and, in addition, the type of function necessary to produce the sidelobe reduction. Now, in order to determine exactly which sidelobes depend upon this section, it is obvious that the aforementioned function should be applied to the original frequency spectrum followed by the evaluation of the corresponding time signal via use of the Inverse Fast Fourier Transform (IFFT). The result of this procedure indicates that only the near-in sidelobes still exist. However, the preceeding procedures have resulted in the first stage of the development of a sidelobe suppression filter.

Having only the near-in sidelobes remaining suggests that the mainlobe be extended to include these sidelobes.

This alteration can be achieved by increasing the amplitude of the signal that lies between the present main-lobe and the near-in sidelobes such that the near-in sidelobes now appear as part of the new main-lobe. The sepctrum which results from this alteration now possesses some components at the band edges which require some reduction in order to satisfy the system bandwidth requirements. However, the results of the second stage of the development process clearly indicates that the desired filter is approaching completion.

The preceeding requirement can easily be satisfied by a multiplicative exponential term with a negative exponent. Applying this function to the latest spectrum yields a time function with a reduced peak amplitude and with range sidelobes that are -34 dB below the main-lobe. This reduction in peak amplitude which produces a SNR loss of approximately 1 dB could have been predicted beforehand because it is obvious that this weighting function reduces the total amount of energy possessed by the signal. This loss in sidelobe level, although undesirable, does not impair system operation at all.

The resultant sidelobe suppression filter is presented in Fig. 6. The most striking aspect of the result is the zero phase angle required by the filter. This fortuitous result allows the sidelobe suppression filter to be combined with the de-chirp filter to yield a single filter whose amplitude characteristics are those of the sidelobe suppression filter and whose phase characteristics are those of the de-chirp filter.

Another significant item about this filter is the resemblence between it and the Taylor filter which has been shown to be the solution of the sidelobe suppression problem when time gating is not employed.

Using the aforementioned techniques, a family of these filters has been derived and Fig. 7 displays a comparison of the resultant sidelobe levels. The results of this derivation indicated that the filter structure approximated the Taylor Filter very closely for large Dispersion Factors. This is an expected result because the time gating effect diminishes as Dispersion Factors increases.

The conclusions of Fig. 7 can be summarized as follows:

1. Significant distortions and, thus higher sidelobes exist for small D.
2. Lower sidelobes can be achieved for larger bandwidths.
3. For a given Dispersion Factor, lower sidelobes can be achieved with increasing bandwidth because less spectral shaping occurs.

The technique used for derivation of the filter can be summarized as follows:

1. Divide the time and frequency signals into desirable and undesirable regions.

2. Perform a convenient alteration on the original time signal and determine filter required for such changes.

3. Eliminate all non-realizable changes and apply remaining components to the original signal. Then, eliminate all components outside the bandwidth.

It should be pointed out that the preceding procedure is most effectively employed in situations where an analytical analysis cannot be performed.

Since the procedure makes use of programs employing the FFT and the IFFT, it is possible to eliminate either one by use of proper normalizing functions.

SUMMARY AND CONCLUSIONS

This paper analyzes and solves the range sidelobe problem that is encountered in all chirp radar systems that employ time gating of "passively generated" chirp pulses. The time gating being employed to produce realizable (finite width) pulses.

The paper presents the derivation of the hitherto unknown chirp receiver response to the time gated transmitted pulses which are "passively generated". Since the results of the derivation clearly indicate that the receiver response has an F-M structure. One is now able to understand the reason for the failure of the paired-echo theory to achieve range sidelobe reductions.

The paper then goes on to describe a method for the derivation of a filter which achieves range sidelobe suppression and then presents a comparison of a family of these filters. One significant result of this derivation lies in the fact that the sidelobe suppression filter can be easily combined with the de-chirp filter, because of its zero phase angle, to yield a single filter whose amplitude characteristics are those of the sidelobe suppression filter and whose phase characteristics are those of the de-chirp filter.

From the family of filters it is easy to conclude that as the effect of the time gating diminishes, the required filter approaches the Taylor filter. Figure 7 in addition displays the relationship between chirp systems having equal Dispersion Factors any varying bandwidth. The figure clearly indicates, as one would expect, that lower sidelobes can be achieved with larger bandwidths and that this effect is more pronounced for small Dispersion Factors.

REFERENCES

1. Klauder, J. R., Price, A. C., Darlington, S., Albersheim, W. J., "The Theory and Design of Chirp Radars," Vol. XXXIX, No. 4, July 1960.

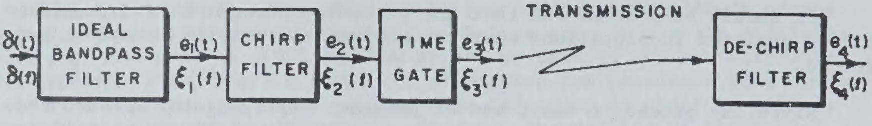

TRANSMISSION
$\delta(t)$
$\delta(f)$
IDEAL BANDPASS FILTER
$e_1(t)$
$\xi_1(f)$
CHIRP FILTER
$e_2(t)$
$\xi_2(f)$
TIME GATE
$e_3(t)$
$\xi_3(f)$
DE-CHIRP FILTER
$e_4(t)$
$\xi_4(f)$

FIG. 1 TIME GATED CHIRP RADAR SYSTEM

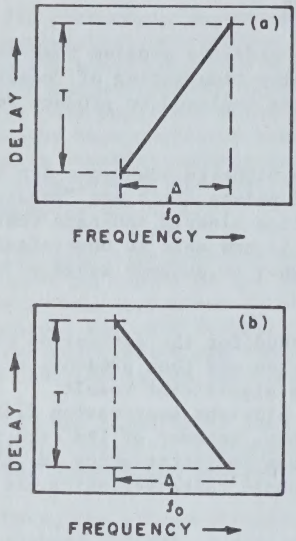


FIG. 2 CHARACTERISTICS OF THE CHIRP AND DE-CHIRP FILTERS, RESPECTIVELY

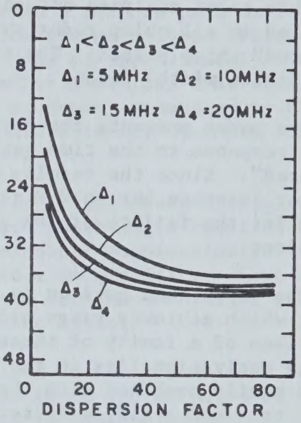


FIG. 7 SIDELOBE LEVEL (BELOW MAIN LOBE) vs. DISPERSION FACTOR.

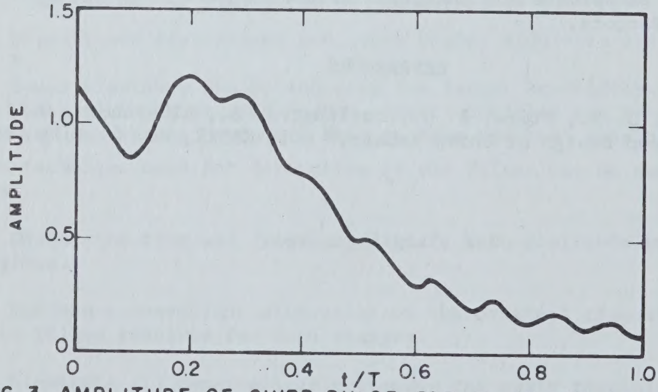


FIG. 3 AMPLITUDE OF CHIRP PULSE FOR D = 10.125. SHAPE IS SYMMETRIC ABOUT THE POINT t/T = 0.

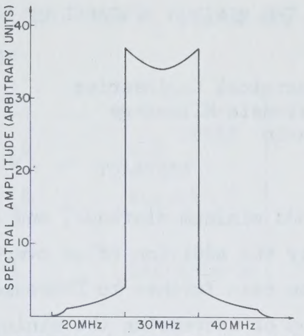


FIG. 5 DE-CHIRP OUTPUT SPECTRUM. PHASE ANGLE IS LINEAR.

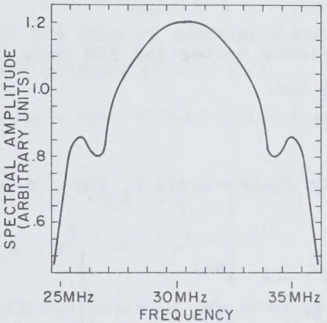


FIG. 6 AMPLITUDE SPECTRA OF THE SIDELOBE SUPPRESSION. PHASE SPECTRA EQUALS ZERO.

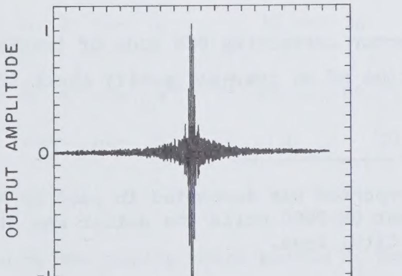


FIG. 4 THE DE-CHIRP FILTER OUTPUT

MULTIPLE EXTENSION OF THE MINIMUM DISTANCE OF BCH CODES*

GEORGE I. DAVIDA
Department of Electrical Engineering
University of Wisconsin-Milwaukee
Milwaukee, Wisconsin 53201

ABSTRACT

Given a code of odd minimum distance, one can increase the minimum distance to $d_{odd} + 1$ by the addition of an over-all parity check [1]. This idea was taken one step further to increase the minimum distance by one more [2]. A method of increasing the minimum distance by two more is presented.

I. INTRODUCTION

Let $\alpha \varepsilon GF(2^m)$. A t-error correcting BCH code is a cyclic code such that every codeword f(x) has

$$\alpha, \alpha^3 \text{ --- } \alpha^{2t-1}$$

as its roots. The parity check matrix H_1 for a t-error correcting BCH code is

$$H_1 = \begin{bmatrix} 1 & \alpha & \alpha^2 & \text{---} & \alpha^{n-1} \\ 1 & \alpha^3 & \alpha^6 & \text{---} & \alpha^{3(n-1)} \\ \vdots & & & & \\ 1 & \alpha^{2t-1} & & \text{---} & \alpha^{(2t-1)(n-1)} \end{bmatrix} \tag{1}$$

II. EXTENDED CODES

Consider the t-error correcting BCH code of length n and distance $d_{odd} = 2t + 1$. Addition of an over-all parity check increases the length

---

* The research reported was supported in part by National Science Foundation Grant GK-2990 while the author was at the University of Iowa, Iowa City, Iowa.

and minimum distance by one [1]. Let H_2 denote the parity check matrix of the extended BCH code, where

$$H_2 = \begin{bmatrix} 1 & 1 & 1 & -- & --- & 1 \\ 0 & 1 & \alpha & \alpha^2 & --- & \alpha^{n-1} \\ 0 & 1 & \alpha^3 & \alpha^6 & --- & \alpha^{3(n-1)} \\ \vdots & & & & & \\ 0 & 1 & \alpha^{2t-1} & & --- & \alpha^{(2t-1)(n-1)} \end{bmatrix} . \tag{2}$$

Adryanov and Saskovets were able to increase the minimum distance by one more by annexing more check digits as follows:

$$H_3 = \left[\begin{array}{ccccc:ccccc} 1 & \beta & \beta^2 & --- & \beta^{n-1} & 1 & \beta & \beta^2 & --- & \beta^{b-1} \\ \hdashline \multicolumn{5}{c:}{H_2} & & & 00 & --- & 00 \end{array}\right] \tag{3}$$

where H_2 is the extended BCH code of minimum distance d, $\beta = \alpha^{d-1}$ and b is the degree of the minimum polynomial of β [2].

Consider the distance - four extended Hamming code with the parity check matrix

$$H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 & --- & 1 \\ 0 & 1 & \alpha & \alpha^2 & --- & \alpha^{n-1} \end{bmatrix} . \tag{4}$$

This code can detect all patterns of three or less errors. Let

$$H_5 = \begin{bmatrix} 0 & 1 & \alpha & \alpha^2 & --- & \alpha^{n-1} & & & \\ 1 & 1 & 1 & 1 & --- & 1 & & & \\ 0 & 1 & \alpha^3 & \alpha^6 & --- & \alpha^{3(n-1)} & I_{mxm} & & \\ 1 & 1 & 1 & 1 & --- & 1 & 11 --- 11 & & \\ 0 & 1 & \alpha^5 & \alpha^{10} & --- & \alpha^{5(n-1)} & 00 --- 00 & I_{mxm} & \\ 0 & 0 & --- & & --- & 0 & I_{mxm} & 00 --- 00 & I_{mxm} \end{bmatrix} \tag{5}$$

Theorem

The codes given by the parity check matrix H_5 have minimum distance of 7.

Proof

Case 1: The sum of some combination of four columns from the first $n + 1$ positions may null out in the first and third blocks of rows, that is, in the α's and fifth powers of α, and be non-zero in the third powers of α. Addition of any two columns from positions $n + 1$ to $n + m$ would require two columns from $n + 2m + 1$ to $n + 3m$.

Case 2: We know that the first two blocks of rows consist of a distance 8 code. We know from BCH construction that the sum of any six from the first n columns will be non-zero in the last block of rows. Thus, at least one more column from positions $n + 2m + 1$ to $n + 3m$ will be needed.

The above construction method can be applied to any BCH code to increase the minimum distance by four.

III. REMARKS

The above codes are puncturable, that is, the additional parity checks can be removed to get the original code. Puncturable codes are interesting if one considers a feedback communication strategy where one first detects errors and, upon detection, requests additional redundancy to correct those errors [3].

References

1. Peterson, W. W., Error-Correcting Codes, MIT Press, 1961.
2. Berlekamp, E. R., Algebraic Coding Theory, McGraw-Hill Inc., N. Y., 1968.
3. Davida, G. I., "Forward Error Correction with Decision Feedback", Ph.D. Thesis, Department of Electrical Engineering, University of Iowa, Iowa City, Iowa.

ON SHORTENED REED-MULLER CODES

C. L. CHEN
Coordinated Science Laboratory, University of Illinois, Urbana, Ill.
S. LIN
Department of Electrical Engineering, University of Hawaii,
Honolulu, Hawaii

Abstract - It is shown that the Reed-Muller codes can be shortened so that the number of parity check digits is reduced. The shortened codes are still majority-logic decodable with the same error-correcting capability. These shortened codes can be used to construct a class of majority-logic decodable convolutional codes in a similar fashion as the Reed-Muller-like codes by Massey.

Let v_0 be an all 1's n-tuple, where $n = 2^m$, G be an $m \times n$ matrix whose columns consist of all possible m-tuples over GF(2). Denote the row vectors of G by $v_1, v_2, \ldots, v_m$. Then the r-th order Reed-Muller (RM) code is formed by using as a basis the vectors $v_0, v_1, \ldots, v_m$, and all vector products of these vectors r or fewer at a time. The code length n, dimension K, and minimum distance d are equal to

$$n = 2^m \qquad K = \sum_{i=0}^{r} \binom{m}{i}$$
$$d = 2^{m-r}$$

Now suppose that we delete from the generator matrix of the r-th order RM code the submatrix formed by the vectors v_0, v_1, $v_2, \ldots, v_s$, and all vector products of these vectors r or fewer at a time, where $s < m$. Then the row rank of the new matrix is equal to

$$K' = K - \sum_{i=0}^{r} \binom{s}{i}$$

where the convention $\binom{s}{i} = 0$ for $i > s$ is used. The new matrix thus formed consists of 2^s all 0's columns. These 2^s all 0's columns can be deleted without reducing the weight of a vector in the row space of the new matrix. Therefore the shortened code generated by the new matrix with 2^s columns deleted has the following parameters

$$n' = 2^m - 2^s$$
$$K' = K - \sum_{i=0}^{r} \binom{s}{i}$$
$$d' = d = 2^{m-r}$$
$$n' - K' = n - K - \sum_{i=r+1}^{s} \binom{s}{i}$$

Notice that the shortened code has the same minimum distance as the original code while the number of parity check digits is reduced by $\sum_{i=r+1}^{s} \binom{s}{i}$.

It is well-known that RM codes are majority-logic decodable.(1,2) From the way we deleted the columns and rows from the generator of RM codes, it is easy to see that the same majority-logic decoding scheme applies to the shortened codes. That is, the shortened codes are also majority-logic decodable.

The RM codes with the over-all parity check digit deleted and with some rearrangement of the positions of digits have been shown to be cyclic.(3) Thus the shortened RM codes are shortened cyclic codes. This fact can be used for encoding the shortened codes.

The shortened RM codes can be used to construct a class of majority-logic decodable convolutional codes in a similar fashion as the Reed-Muller-like codes by Massey.(4) Each of the columns in the generator matrix of a shortened code is the last row of a parity triangle. The parameters of the codes thus formed are

$$\text{rate} = \frac{1}{n'+1}$$

$$\text{constraint length} = (n'+1)k'$$

$$\text{minimum distance} = \sum_{i=1}^{r} 2^{m-i}\left[\binom{m}{i}-\binom{s}{i}\right] + 1$$

REFERENCES

1. Reed, I. S., "A Class of Multiple-Error-Correcting Codes and the Decoding Scheme," IRE Trans., IT-4, 34-49 (1954).

2. Peterson, W. W., Error-Correcting Codes, MIT Press, Cambridge, Mass. (1961), 73-77.

3. Kasami, T., S. Lin, and W. W. Peterson, "New Generalizations of the Reed-Muller Codes - Part I: Primitive Codes," IEEE Trans., IT-14, 189-199 (1968).

4. Massey, J. L., Threshold Decoding, MIT Press, Cambridge, Mass. (1963), 49-51.

PROBABILISTIC GENERATION, TRANSMISSION AND SYNTACTICAL DECODING OF CONTEXT-FREE PROGRAMMED LANGUAGES

C.R. Souza
University of Notre Dame
Notre Dame, Indiana

and

R.A. Scholtz
University of Southern California
Los Angeles, California

ABSTRACT

It is proposed that syntactical information be used to aid in the post-information-theoretic decoding of formal languages that have been transmitted through a noisy channel. Context-free programmed grammars with probabilities attached to the "go-to" fields are studied as realistic models for syntactical information sources. The model is formally defined and an example of its output is given. Simplifications that follow from imposing leftmost derivations are displayed. Source models for languages that are known to be context-free are studied, and for these a first-order approximating source is obtained; standard methods may then be used to calculate its eutropy. General desirable properties of syntactical decoders are proposed, and a general minimum-distance algorithm established. Backtracking s-grammars are defined and used to exemplify backtracking algorithms for detection and correction of single and multiple errors.

INTRODUCTION

The usefulness of a new theory can be measured not only in terms of its intrinsic success in the description and prediction of the, natural or mathematical, phenomena it treats, but also in terms of its ability to generalize, improve on, extend and solve the problems of, previously developed theories in the same or in different fields.

As it is well known, the present theory of Formal Languages was born from efforts to model the syntactical structure of natural languages, and received a great boost when its ability to model some programming languages was discovered [1]. It led naturally to among others, studies of the translation process performed by compilers [2] and to formalizations of semantic evaluations [3] and of still better models for the syntax of natural languages [4], not to mention its effect in the study of devices, Automata Theory [5]. On the other hand, Information Theory was born from efforts to model the process of communication [6], achieved prominence through its impact on communications technology and was applied widely to the most varied fields, though falling short of an acceptable description for the process of transference of ideas or commands that occurs when intelligent beings communicate among themselves and with machines [7,8]. Efforts to connect these two, so clearly related, theories have been sporadic [9,10], but recently Probability Theory and the theory of Formal Languages are increasingly interacting [11,12] directly, rather than through probabilistic devices as in [13].

This paper describes the work done by the authors on the problem of modelling the noisy transmission of a formal language through a communication channel and the exploitation of the syntactical information embedded in the language to aid in the detection, location and correction of source or trasmission errors.

A complete model for the generation of messages or sentences, their transmission and reception is displayed in Figure 1. The ordering "Semantics — Syntax — Lexicon" is somewhat arbitrary and not intended to model exactly the process by which human beings, say, generate and receive sentences. The inner dotted block represents the classical transmission system model of Information Theory, the end blocks being usually blurred together and lebelled simply "Source" and "User".

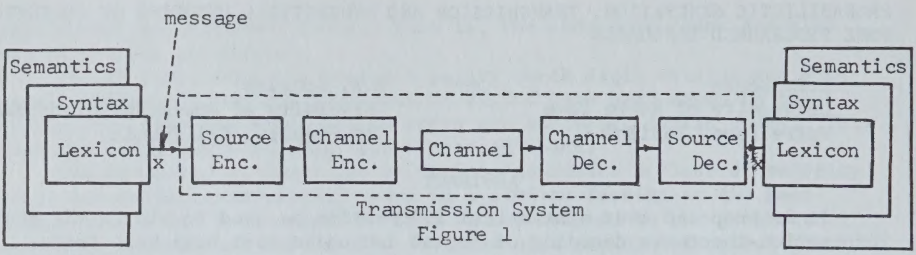

message
Semantics
Syntax
Lexicon
x
Source Enc.
Channel Enc.
Channel
Channel Dec.
Source Dec.
x
Semantics
Syntax
Lexicon
Transmission System

Figure 1

Since the formalism to handle syntax is so much more developed than that to handle semantics, the authors restricted their attention to the model of Figure 2, where the inner dotted block is simply labelled "Transmission System"; the model for the source was taken as a context-free programmed grammar with probabilities, cfpgp, following a suggestion of P. Dierks, [16], consisting of a context-free programmed grammar as in 14 , with probability distributions assigned to both success and failure fields. It was studied formally and by simulation.

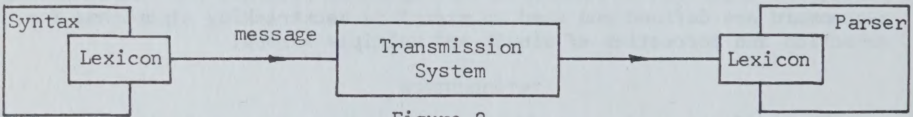

Syntax
Lexicon
message
Transmission System
Parser
Lexicon

Figure 2

The simulating program was written in LISP for the IBM 360 and a large number of grammars and probability assignments were tried. Several of the grammars and the resulting outputs were shown at the Conference. One simple such grammar is shown in Figure 3 and a small portion of the resulting output in Figure 4.

```
SETUP (
      ((*  25))
(
(1  (S   *)                (2 3 4 5 )      (0.27 0.54 0.81 1.0))
(2  (*    *)               (2 3 4 5 )      (0 05 0.40 0.81 1.0))
(3  (*  < * - * $ )        (2 3 4 5)       (0.14 0.28 0.42 1.0))
(4  (*  < * % * $ )        ( 2 3 4 5 )     (0.14 0.28 0.42 1.0 ))
(5  (*    )                (6)             (1.0))
(6  (-  +)      (2 3 4 5) (0.14 0.28 0.42 1.0) (7 8) (0.5 1.0))
(7  (%  &)      (2 3 4 5) (o.14 0.28 0.42 1.0) (8) (1.0))
(7  ($  >)                (6) (1.0) (7) (1.0))
```

Figure 3

```
(< E & R >)
(V)
(  < < H &   < N + L > > + W >)
(  < C + I >)
(Y)
(< D + M >)
(< R + < < < K + < C + Y > > + A > & M > >)
(  < < < M & J > & Z > + C >)
(< B +   < < < M + < T + E > > & A > + < X + E > > >)
(  <   < K + W > + N >)
(< Z + P >)
(T)
```

Figure 4

In order to implement a desired scheme, similar to that of Sequential Decoding [17], the external statistical properties of the source had to be better known. It was clear that the typical sequences of failures and "hang-ups" of a context-free programmed grammar made the relation between the internal workings of the source (and the statistical properties of the production label sequences) and the resulting output sequence of terminal symbols mathematically intractable. Even though the label sequences of cfpgp are very similar to the state sequences of a probabilistic automaton [18], it is easy to show that the autonomous behavior of the cfpgp in general, is not, Markovian of any order [19]; that is because of the undeveloped string provides a potentially infinite memory, its length not being restricted, and hence present transitions may depend on labels as far in the past as one wishes. Finite-order Markovian approximations then seem desirable.

We shall show firstly that, given any cfpgp, an equivalent cfpgp can be constructed in such a manner that it is "fail-safe", if leftmost derivations are imposed. For basic definitions and nomenclature, the reader is referred to [14].

SOURCE MODEL

Definition 1: A cfpgp is a context-free programmed grammar $G = <N,T,P,S,J,F>$ with a probability distribution function associated with each "go-to" field. The typical form of a production is

$$i)\ A_i \rightarrow \psi_i \quad S(V_i)P(V_i)F(W_i)P(W_i)$$

where if $(V_i) = (j,k,l, \ldots)CJ$, then $P(V_i) = (P^S_{ij}, P^S_{ik}, P^S_{il}, \ldots)$, where

$$P^S_{ij} = \text{Pr. }\{j \text{ is chosen}/\ i \text{ was successfully applied}\}$$

is a conditional probability, etc. Similarly, $P(W_i)$ will be composed of probabilities P^F_{ij}.

Let us examine for a moment what could happen after the application of production j) if the leftmost nonterminal symbol in the intermediate string is always the first to be processed: if the right-hand side of the production core has at least one nonterminal, let the leftmost of them be denoted by B. It is known, then, that B must be processed next, since it must be replacing the previous leftmost terminal, A_j. Searching V_j, we may find B-rules; these should remain in V_j. Those labels in V_j which do not correspond to B-rules are of the form

$$i)\ A_i \rightarrow \psi_i \quad S(V_i)P(V_i)\ F(W_i)P(W_i),\ A_i \neq B;$$

and we, therefore, can search through all such W_i for B-rules, repeating the search through failure fields until all reachable B-rules are found. Their labels can then be inserted in V_j. The probabilities of the corresponding paths can be calculated even when there are failure-field loops [15] and inserted in $P(V_j)$.

Repeated application of the above procedure would eliminate all failure fields for productions reachable from such productions j).

Suppose now that the right-hand side of production j) does not contain nonterminals; then the productions in $S(V_j)$ may or may not be applicable, depending on which nonterminal is next in the undeveloped string, if any (the string could also have been fully developed). By allowing a single nonterminal look-ahead, we can build fields conditioned in the appearance of each possible nonterminal, and compute the appropriate path probabilities. We have then:

Theoreml. Any cfpgp with leftmost derivations can be put in an equivalent form (also equivalent statistically) with no failure fields and a single nonterminal look-ahead.

Proof: By repeated application of the procedures outlined above, the successful application of any production that may be reached from the successful application of any other production is guaranteed. Since the initial application of an S-rule is guaranteed to succeed, no production will ever fail and therefore all failure fields will then be useless ∎

The typical form of the productions can then be simplified to

$$j)\ A_j \rightarrow \psi_j \quad (V_j)P(V_j) \quad \text{if } \psi_j \in V^*NV^*$$

or

$$j)\ A_j \rightarrow x \quad A{:}(V_j^A)P(V_j^A) \quad B{:}(V_j^B)P(V_j^B) \ \ldots \quad A,B,\ldots \in N,\ x \in T^*.$$

Example 1: The following grammar generates boolean forms with x standing for any variable:

1) $S \rightarrow \sim S$ S(5) (1) F(6) (1)
2) $S \rightarrow (SBSA$ S(1,2,3,4) (1/4,1/4,1/4,1/4) F(6) (1)
3) $S \rightarrow (SCSA$ S(1,2,3,4) (1/4,1/4,1/4,1/4) F(6) (1)
4) $S \rightarrow x$ S(5,6) (1/2,1/2) F(6) (1)
5) $B \rightarrow V$ S(1,2,3,4) (1/4,1/4,1/4,1/4) F(1,2,3,4,6) (1/5,1/5,1/5,1/5,1/5)
6) $C \rightarrow \Lambda$ S(1,2,3,4) (1/4,1/4,1/4,1/4) F(5,7) (1/2,1/2)
7) $A \rightarrow)$ S(5,6) (1/2,1/2) F(6) (1)

Applying the procedure above, the following equivalent grammar is found:

1) $S \rightarrow \sim S$ (1,2,3,4) (1/4,1/4,1/4,1/4)
2) $S \rightarrow (SBSA$ (1,2,3,4) (1/4,1/4,1/4,1/4)
3) $S \rightarrow (SCSA$ (1,2,3,4) (1/4,1/4,1/4,1/4)
4) $S \rightarrow x$ S: (1,2,3,4) (1/4,1/4,1/4,1/4) A: (7)(1) B: (5)(1) C: (6)(1)
5) $B \rightarrow V$ S: (1,2,3,4) (1/4,1/4,1/4,1/4) A: (7)(1) B: (5)(1) C: (6)(1)
6) $C \rightarrow \Lambda$ S: (1,2,3,4) (1/4,1/4,1/4,1/4) A: (7)(1) B: (5)(1) C: (6)(1)
7) $A \rightarrow)$ (1,2,3,4) (1/4,1/4,1/4,1/4) A: (7)(1) B: (6)(1) C: (6)(1)

where all failure fields have been eliminated.

If the language being considered is known to be context-free (which is not always the case with programmed grammar with context-free cores) then we may as well start with a standard-form grammar for it [20] and adopt the following model:

Definition 2. A "standard context-free programmed grammar with probabilities," spgp, is a programmed grammar with productions of either of two forms:

$$j)\ A_j \rightarrow a_j Y_j^i \ldots Y_j^n \quad (V_j)P(V_j) \quad a_j \in T,\ A_j \in N,\ (Y_j^1 \ldots Y_j^n) \in N^+$$

where V_j contains the labels of all Y_j^1 rules or

$$j)\ A_j \rightarrow a_j \quad A{:}\ (V_j^A)P(V_j^A) \quad B{:}\ (V_j^B)P(V_j^B),\ \ldots,\ a_j \in T;\ A_j, A, B, \ldots \in N$$

where V_j^A contains the labels of all A-rules, etc.

An spgp always produces a terminal, or output symbol, whenever a production is applied; the relationship between the label sequence and the output sequence is a simple map, very easy to describe and treat mathematically and allows one to derive interesting approximations concerning stationary distributions, entropy, etc. of the language. To simplify the calculation of probabilities we introduce a restriction in the $P(V_j)$:

Definition 3. An spgp is said to have a "consistent probability assignment" iff for every $B \in N$, whenever V_j or, for some $A \in N$, V_j^A, contain all the B-rule labels, then the $P(V_j)$ and $P(V_j^A)$ are the same.

Spgp's with consistent probabilities are such that the probability of a certain B-rule label, say i), be chosen among the other B-rule labels is the same regardless of the field in the grammar from which the control was transferred to the B-rules. We will denote these probabilities by p_{-i}.

Definition 4. A string $Z \in N^*$ is a "successor of A," $A \in N$, written ZsA, iff some intermediate string of the form xVAZW, $x \in T^+, V, W, \in N^*$, can be derived from S. Z will be said a "successor of A with respect to production i) whenever the time to process A comes.

Zs_iA is actually an event, for which we will calculate a probability.

Our model is now a short step from a first-order Markov approximation of the workings of the label chain. Indeed, productions of the form

$$\text{i)}\quad A_i \rightarrow a_i Y_i^j \ldots Y_i^n \quad (V_i)P(V_i)$$

provide stationary transfer probabilities $[p_{ij}]$ for the production labels, very much as those of a first-order Markov chain. The problem resides in the "erasing" productions

$$\text{j)}\quad A_j \rightarrow a_j \quad A: (V_j^A)P(V_j^A)\ldots$$

in which the transfer probabilities are conditioned on the occurrence of a given nonterminal ahead of A_j in the undeveloped string. This occurrence can be ascertained with probability 1 by observing all the past label sequence from the application of the first S-rule to the current application of j); clearly this sequence can be arbitrarily long.

The source model, much like a chatty housewife, is to produce a continuous stream of sentences. This is easily implemented in spgp's by introducing a terminal endmarker, say ¢, and a nonterminal endmarker, say \$, and a new production

$$\ell)\quad \$ \rightarrow ¢S\$ \quad (V_\ell)P(V_\ell)$$

where S is the former initial symbol and (V_ℓ) contains all S-rules, \$ is treated as any nonterminal but ℓ) is the only \$-rule and should be used as the very first rule to be applied; that is, \$ can be considered as a new initial symbol. The grammar will never generate a terminal string since \$ cannot be erased, but the intermediate strings will appear as

$$¢x¢y¢\ldots¢vZ\$$$

where x,y,..., are sentences in the language, $v \in T^*$, $Z \in N^*$.

Let us assume that steady-state probabilities exist for an approximating first-order chain. We may attempt to compute the corresponding approximation to Pr {BsA} for all A,B. Let:

$$p = (p_1, p_2, \ldots p_r) \qquad J = \{1,2,\ldots r\}$$

represent the steady-state distribution row vector for labels 1,2,...r. The following theorem may be stated:

Theorem 2. Assuming that p exists, it is possible to calculate the probability that, given that A_k has a successor with respect to k), this successor is B. The calculation can be effectively performed by the following steps: for all A_k,B,

1) list all productions i) in which A_kB appears in the right hand side*. The occurrence of this set of labels will be denoted as the event $\{Bs^\circ A_k\}$. Its probability is the sum of the steady-state probabilities of occurrence of all the labels in the set. Let $\Pr\{Bs_k^1A_k\} = \Pr\{Bs^\circ A_k\}p_{-k}$.

2) list all productions j), as in Theorem 4.1, in which A is the last nonterminal. Let $\Pr\{Bs_k^2A_k\} = \Pr\{Bs_k^1A_k\} + \sum_j \Pr\{Bs_j^1A_j\}p_{-k}$; noting that $\Pr\{Bs_j^1A_j\}$ has been computed in step 1). Let: $\Pr\{Bs_k^nA_k\} = \Pr\{Bs_k^{(n-1)}A_j\}p_{-k}$.

Then, if for some n, $\Pr\{Bs_k^nA_k\} = \Pr\{Bs_k^{(n+1)}A_k\}$ set

---

* If some production contains several pairs A_kB, its label should be counted the same number of times.

$\Pr\{Bs_k A_k\} = \Pr\{Bs_k^n A_k\}$. Otherwise let $\Pr\{Bs_k A_k\} = \lim_{n\to\infty} \Pr\{Bs_k^n A_k\}$, and

$$\Pr\{Bs_k A_k / \exists Ds_k A_k\} = \frac{\Pr\{Bs_k A_k\}}{\sum_{D\varepsilon N} \Pr\{Ds_k A_k\}} .$$

Proof: See [15].

With approximating probabilities for the events $\{Bs_i A_i\}$ computed as above, transfer probabilities $\{p_{ij}\}$ may now be calculated for all producttions; those corresponding to "erasing" productions will be, in general, functions of p. Therefore the transition matrix for the approximating chain is a function of its own steady-state probabilities.

Theorem 3. Let $\Pi = [p_{ij}] = \Pi(p_i, \ldots, p_r) = \Pi(p)$ be the transition probability matrix, as function of the steady-state distribution p, obtained by the method exposed above. Then if p exists it must satisfy $p\Pi(p) = p$.

Proof: If p exists, by Theorem 2 and its sequel $\Pi(p)$ can be determined as a function of p and the given transfer probabilities of the grammar. It is also known [3] that if p is a steady-state distribution then it is also a stationary distribution for the chain ∎

Off-hand, no statement can be made about the uniqueness of p, since Π is not independent of p. In all cases we have dealt with, the above equation had a unique solution.

Definition 5. Let the uncertainty associated with label j be given by $H_j = H(p_{j1}, \ldots p_{jr})$ where H( ) is the usual average uncertainty measure, or entropy. That is,

$$H_j = - \sum_{i-1}^{r} p_{ji} \log_2 p_{ji} \text{ bits/label.}$$

Theorem 4. The average uncertainty, or entropy, of the label sequence of the first order Markov approximation of a spgp is given by

$$H(J) = \sum_{i-1}^{r} p_j H_j \text{ bits/label.}$$

Proof: The above expression holds for any unifilar Markov source [19]. Since the labels reachable from any j are all distinct, the label chain qualifies as a unifilar source ∎

DECODING

Let us now direct our attention to the other end of the transmission scheme. It is clear that the formal-linguistic model has afforded us a succint description of a (in general infinite) code [19]. The parser provides an equally simple way of checking whether a given sequence of symbols is a code word or not, assuming that the chosen language is decidable. We shall be concerned, then, with the problem of utilizing the syntactical information available on the language being transmitted in order to detect, locate and eventually correct errors.

The following seem to be sensible general properties of a well-behaved "syntactical decoder" (S.D.):

1. If a given sentence is received error-free, or if the errors of the T.S. are such that the received sentence is grammatical, then such sentence should be forewarded to the user without modifications. Except in "pathological" situations, this is the most probable transmitted sentence. In fact, being "bling" to meaning, a purely syntactical decoder would not be capable of recognizing wrong but grammatical sentences.

2. If the output sentence of the T.S. is not grammatical, the S.D. should attempt to produce a probable grammatical sentence and forward it to the user.

3. Whenever correction is attempted, the user should be so informed: that is, a "flag" should be placed alongside each sentence altered by the

S.D.; additional information about the location and nature of such alterations could, in certain cases, also be wanted. Additional requirements could be that

4. The S.D.'s strategy in 2. above should minimize some measure of computational complexity (time, storage space, number of control states, etc.)

5. There could be a feedback connection from the S.D. to the T.S. decoder to aid in a) stopping error-propagation phenomena and b) changing parameters in adaptive schemes.

There are still at least two possible choices of configuration for the S.D. In the first, the actions of deciding whether the sentences are grammatical and of altering those that are not are clearly separated (see Fig. 5); this may not always be desirable, and the second choice, that of incorporating the transducer and the acceptor, will be treated shortly.

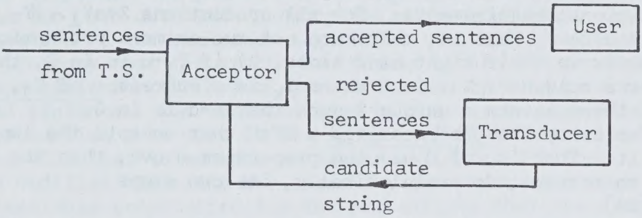


Figure 5

For the first choice, the following procedure will produce a sentence in the language that is "closest" in number of substitutions to the received sentence; we will word it for deterministic, one-way acceptors:

Procedure 1.

Step 1. Given the position in the sentence where the parser "hang-up" replace that symbol with another from T and return the sentence to the parser until T is exhausted.

Step 2. If the above steps do not produce an acceptable sentence, return the original received symbol to the position, move left one symbol and again run through T.

Step 3. Repeat (2) until either the beginning marker is reached or the sentence is accepted. If the marker is reached, return to the "hang-up" point and start substituting pairs of symbols, then triples, etc.

From a computational time point of view, however, the above algorithm has serious shortcomings: The first is that the sentence must be parsed again from the very beginning every time a substitution is made; this problem is solved below for a particular class of languages. The second, particularly important for languages with large sets T, is that syntactical information is not used in picking a likely symbol to try; this problem is still very much open.

In order to save space the definitions of s-language and s-machine are omitted; the reader is referred to [20], [21], [22] and [23] for the necessary concepts and notation.

Definition 6: A Back-Tracking s-Machine, BTsM, is a two-state pushdown automaton with the following properties: 1) in its first state, its input head moves from left to right only, and the BTsM behaves exactly as an s-machine; and 2) in the second state its input head moves right to left, and it can erase any number of symbols from the top of the push-down tape, proceeding in exactly the reversed manner of its forward moves.

The instructions for the BTsM can be written in the following manner:

$$(s_0,a,Z) \to (s_0,Y_1\ldots Y_n, +1)$$

$$(s_0,a,Y_1) \to (s_i,b^n,Z,\pm 1) \qquad s_i \in \{s_0,s_1\}$$

the first being a standard s-machine instruction and the second containing the number of erasures to be made (or "blanks," b, to be written) before printing z and moving the input head.

It is necessary, then, to find the conditions under which, given an s-grammar, a deterministic BTsM can be found for it; that is, the properties of the s-grammar which allow its parser to backtrack deterministically.* An auxiliary Definition is needed here:

Definition 7: Given an s-Grammar G, a nonterminal Y_k is a "successor" of Y_j iff 1) there is some production $Z \to a \ldots Y_j Y_k \ldots$ or 2) there is a production $Z \to a \ldots Y_j$ and Y_k is a successor of Z.

The reader should forgive us a definition which may be recursive; the number of productions and nonterminals being finite, and consequently the number of last symbols being finite, it is always possible to check whether Y_k is a successor of Y_j or not.

Definition 8: A "Back-Tracking s-Grammar,"BTsG, is an s-grammar with the following additional requirements: for all productions $Z \to aY_1 \ldots Y_n$, 1) the pair (Y_1, a) is unique; that is, there are not two or more productions with the same 2-prefix in their right hand sides. 2) if $Y_j \to c$ is in G, then there does not exist a rule $W \to cY_k \ldots$ in G where Y_k is a successor of Y_j, and 3) for every $c \to T$ there exists a unique Z such that $Z \to c$ is in G.

Theorem 5. For every BTsG there exists a BTsM that accepts the language generated by it. That is, if G has the properties above, then the s-machine will be able to retrace, deterministically, its own steps.

Proof: See [23].

BTsM's can then be used to improve Procedure 1. Let us consider first the procedure that may be used to attempt the correction of a single detectable error per sentence, a detectable error being one which transduces a sentence in the language to another not in it. The following steps are guaranteed to give a sentence in the language at distance 1 from the received one, if one exists:

Procedure 2.

Step 1. Cycle through T at the hang-up point, attempting to continue the forward parsing. If the parsing of the sentence cannot be continued, always return, backtracking to the hang-up point to try another symbol.

Step 2. The above attempts failing, move one place left from the hang-up point and cycle through T, attempting again the forward parse.

Step 3. If the symbol ¢ is reached backtracking, there is no sentence in the language at distance 1 from the one received.

It is clear that to do this the BTsM must have some additional storage capacity: namely, it must know where the substitutions were attempted and whether T has been exhausted or not. This can be done nicely with a double-track input tape: the first track, containing the input string and the second, auxiliary track, storing the necessary information. This type of machine will be called "Double-Track BTsM," DBTM.

Theorem 6. For every BTsG there exists a DBTM which accepts the language generated by it and, when presented with a sentence not in the language, will deterministically produce one sentence in the language at distance 1 from the given one, provided one exists.

Proof: See [23].

A generalization of Procedure 2 will allow the DBTM to search for two, three and multiple error patterns up to any fixed finite number N of errors per sentence. We may state

Theorem 7. For every BTsG there exists a DBTM which accepts the language generated by it and, when presented with a sentence not in the language, will deterministically produce one sentence in the language the fewest possible number of substitutions away from the given sentence, provided one exists within a pre-established maximum number of substitutions, N.

---

* These restrictions are not necessary if the machine is allowed to print, along side its input tape, the number of symbols pushed down and the symbol erased. This information then can be used for backtracking.

Proof: See [23]. In [23] complete descriptions of the single-error-correcting DBTM and of the multiple-error-correcting DBTM are given. (We are using "correction" in the sense of "attempted correction"). They have been successfully simulated in SNOBOL.

For general cfpgp's, under reasonably general conditions, an auxiliary memory in the form of a two-way read-write tape with length proportional to that of the received sentence is sufficient additional storage to provide backtrack capability.

FINAL COMMENTS

The virtue of backtrack programming is that it is an efficient and exhaustive algorithm for searching trees. Sometimes even this is not enough to reduce computational problems to a tractable size. In this case the main hope for both automatic parsing and error-correction appears to lie in a modification of the backtrack algorithm. The basic algorithm stated above has the property that the system backtracks within the production tree when it is impossible for the branch just traversed to lead to acceptance of the estimated sentence. The modified form of the search algorithm backtracks in the production tree when it is highly unlikely that the branch just traversed will lead to acceptance of the estimated sentence.

The modified algorithm for searching the production tree would require an implementable computation for the probability that the present branch will lead to an acceptance of the estimated message. Actual algorithms for carrying out this type of probabilistic search have been studied extensively under the heading of sequential decoding [17]. While the search control program of sequential decoding can be adapted to our tree search problem, the implementation of the probability computation is much more difficult for cfpgp's due to their inherently non-Markovian nature.

From a more general point of view, it can safely be stated that a large share of the traffic of communications in the future will consist of data transmission between users and computers or between computers. Since, aside from purely numerical data flow, quite a portion of this communication is done in one formal language or another (here we are using "formal" in a broad sense) it seem logical to study communications systems tailored for such traffic. The fact that the syntax (and some of the semantics) of these languages is amenable to mathematical treatment seems to suggest that they be used to improve the reliability of the transmission. Some of the modern time-sharing systems already do this, in a sense, by providing line-by-line syntax checks for programs entered from remote stations. We expect syntactical error detection, location and correction to be an increasingly important area of research in the near future.

REFERENCES

1. Aho, A.V. and Ullman, J.D., "The Theory of Languages," Mathematical Systems Theory, Vol. 2, No. 2, pp. 127-146, 1968.
2. Aho, A.V., Hopcroft, J. and Ullman, J.D., "A General Theory of Translation," Mathematical Systems Theory, Vol. 3, pp. 193-221, 1969.
3. Knuth, D., "Semantics of Context-Free Languages," Mathematical Systems Theory, Vol. 2, pp. 127-145, 1968.
4. Ginsburg, S. and Partee, B., "A Mathematical Model of Transformational Grammars," Information and Control, Vol. 15, pp. 297-334, 1969.
5. Hopcroft, J. and Ullman, J.D., Formal Languages and Their Relation to Automata, Addison-Wesley Pub. Co., Inc., Reading, Mass. 1969.
6. Shannon, C.E., "A Mathematical Theory of Communications," Bell Systems Technical Journal, Vol. 27, pp. 379-423, 623-656, 1948.

7. Cherry, C., On Human Communication: A Review, A Survey and A Criticism, M.I.T. Press, 2nd Edition, Cambridge, Mass., 1966.
8. Maiko, H., "Information Theory and Cybernetics," IEEE Spectrum, Vol. 4, No. 11, pp. 75-83, 1967.
9. Conant, R.C., "Channel Capacity of Moore Automata," Information and Control, Vol. 12, No. 5/6, pp. 453, 1968.
10. Banerji, R.B., "Phrase-Structure Languages, Finite Machines and Channel Capacity," Information and Control, Vol. 6, pp. 153-162, 1963.
11. Booth, T.L., "Probabilistic Representation of Formal Languages," IEEE Conf. Rec., 10th Annual Symp. on Switching and Automata Theory, pp. 74-81, Waterloo, Ontario, 1969.
12. Salomaa, A., "Probabilistic and Weighted Grammars," Information and Control, Vol. 15, pp. 529-544, December, 1969.
13. Souza, C.R., "Probabilistic Automata with Monitored Final State Sets," IEEE Trans. on Computers, to appear.
14. Rosenkrantz, D.J., "Programmed Grammars - A New Device for Generating Formal Languages," IEEE Conf. Rec., 8th Annual Symp. on Switching and Automata Theory, pp. 14-20, 1967.
15. Souza, C.R., Probabilities in Context-Free Programmed Grammars," The Aloha System Technical Report A70-2, University of Hawaii, Honolulu, 1970.
16. Dierks, P.H., "Using Grammars to Improve Decoding," EE Tech. Report, University of Notre Dame, Notre Dame, Indiana, in preparation.
17. Wozencraft, J.M. and Reiffen, B., Sequential Decoding, Technology Press, New York, 1961.
18. Souza, C.R. and Leake, R.J., "Relationships Among Distinct Models and Notions of Equivalence for Stochastic Finite-State Systems," IEEE Trans. on Computers, Vol. C-18, No. 7, pp. 633-641, 1969.
19. Ash, R., Information Theory, John Wiley & Sons, New York, 1965.
20. Greibach, S.A., "A New Normal-Form Theorem for Context-Free Phrase Structure Grammars," Journal of the A.C.M., Vol. 12, No. 1, pp. 42-52, 1965.
21. Golomb, S.W. and Baumert, L.D., "Backtrack Programming," Journal of the A.C.M., Vol. 12, No. 4, pp. 516-524, 1965.
22. Korenjak, A.J. and Hopcroft, J.E., "Simple Deterministic Languages," IEEE Conf. Rec., Seventh Annual Symp. on Swithing and Automata Theory, pp. 36-46, 1966.
23. Souza, C.R. and Scholtz, R.A., "Syntactical Decoders and Backtracking s-Grammars," The Aloha System Tech. Report A69-9, University of Hawaii, Honolulu, November 1969.

The research leading to the results presented here was partially supported by the ALOHA SYSTEM, a research project of the University of Hawaii, supported by the Office of Aerospace Research (SRMA) under Contract No. F44620-69-C-0030, a project THEMIS award.

A CLASS OF MAJORITY DECODABLE MULTIPLE SOLID-BURST-ERROR-CORRECTING CYCLIC CODES

HARRY T. HSU
The Pennsylvania State University
University Park, Pa. 16802

ABSTRACT

In this paper, a class of majority decodable solid burst-error correcting cyclic codes will be presented. The advantage of the cyclic codes presented is that the complexity of the decoding algorithm is independent of the multiplicity of the solid-burst-error patterns. For the correction of single and double solid-burst-error patterns, special parity check codes with modified decoding algorithm will be presented so that the number of check digits will be further reduced.

INTRODUCTION AND PRELIMINARIES

Solid-burst error correcting cyclic codes have been studied by several authors i.e. Chien and Tzeng (1969), Hsu (1968), Schillinger (1964), Shiva and Sheng (1969). This paper will present a class of m multiple solid-burst-error-correcting cyclic codes which has a very simple decoding algorithm.

The following notations will be used throughout this paper.

GF(2)	Galois field of two elements
$g(x)$	Code generator polynomial
n	Code length
m	Multiplicity of solid burst error patterns
$B_j(x)$	A single solid burst error pattern of arbitrary length
$B(x)$	$B(x) = \sum_{j=1}^{m} B_j(x)$, the error polynomial
$a(x)$	Code polynomial, $\deg(a(x))=n-2$
$r(x)$	Received code polynomial
c	$(c-1)$ is the number of useful information digits
$\mathrm{maj}(x_1,x_2,x_3,\text{---},x_n)$	majority decision on $x_1,x_2,x_3,$---and x_n
$[x]$	the largest integer $\leq x$
$f(x)=\sum_{i=0}^{(n-1)} f_i x^i$,	where $f_i \in GF(2)$ is the polynomial representation of the n-tuple vector $w=(f_0,f_1,f_2,\text{---},f_{(n-1)})$. Throughout this paper, $f(x)$ and w will be considered as identical.

A solid burst error pattern is defined to be a burst error pattern such that within the burst, all the digits are error digits. A solid burst error pattern of length ℓ can be represented as:

$B(x)=x^j\ (\sum_{i=0}^{\ell-1} x^i)$, where x^j shows the location of the solid burst error pattern. If m multiple solid-burst-error patterns occurred, then the error polynomial $B(x)$ can be represented as:

$B(x)=\sum_{j=1}^{m} B_j(x)$ Since $(1+x)B_i(x)=x^{\ell_i}+x^{j_i}$, where ℓ_i and j_i are artitrary integers, $(1+x)B(x)=\sum_{i=1}^{m}(x^{\ell_i}+x^{j_i})$. $(1+x)B(x)$ contains at most 2m random errors. Therefore any cyclic code with $d_{min} \geq 4m+1$ can be used for the correction of m multiple solid-burst-error patterns.

If $\{a_1,a_2,a_3,---a_{(c-1)}\}$ is the set of information digits, then the decoding algorithm presented in this paper can always lead to the recapture of at least (c-1) digits of the set $\{a_1,a_1+a_2,a_2+a_3,---a_{(c-2)}+a_{(c-1)},a_{(c-1)}\}$. As a result, all the information digits can be computed. This result will be stated in lemma 1.

(Lemma 1) If any (c-1) digits of the set $\{a_1,a_1+a_2, a_2+a_3,---,---,a_{(c-2)}+a_{(c-1)},a_{(c-1)}\}$ can be obtained, then the set of all the information digits i.e., $\{a_1,a_2,a_3,---,a_{(c-1)}\}$ can be computed.

II. PRESENTATION OF RESULTS

In the following a class of majority decodable m multiple solid-burst-error correcting cyclic codes will be presented. Although the rate of transmission is poor, it has the advantage of a very simple decoding algorithm. For single and double solid-burst-error correction, special parity check codes will be presented so that the rate of transmission can be improved.

(A) Cyclic Codes for m Multiple-Solid-Burst-Error Correction

(Theorem 1) $g(x)=\sum_{i=0}^{4m} x^{ic}$ generates a cyclic code of length $n=(4m+1)c$ which is capable of correcting m multiple solid-burst-errors of arbitrary length provided that the actual useful information digits is (c-1).

Since $g(x)=\sum_{i=0}^{4m} x^{ic}$, if $I=(a_1,a_2,a_3,---,a_{(c-1)})$ is the set of information digits, the transmitted code vector w can be expressed as:

$w=(I,0,I,0,---,I,0)$, where there are $(4m+1)(I,0)$ pairs.

Theorem 1 will become obvious when the decoding algorithm is discussed.

The received code polynomial r(x) is:

$r(x)=a(x)+B(x)$, where $B(x)=\sum_{j=1}^{m} B_j(x)$

$(1+x)r(x)=(1+x)a(x)+(1+x)B(x)=(1+x)a(x)+\sum_{i=1}^{m}(x^{\ell_i}+x^{j_i})$, where ℓ_i and j_i are two arbitrary integers. Therefore $(1+x)r(x)$ contains at most 2m multiple random errors. Since a(x) is a code polynomial of degree $\leq (n-2)$, $(1+x)a(x)$ must be a code polynomial of degrees $\leq (n-1)$.

The received code polynomial r(x) multiplied by (1+x) will be the actual received code polynomial used in decoding.

$(1+x)r(x)$ and $(1+x)a(x)$ can be represented as:

$(1+x)r(x)=(r_1,r_2,r_3,---,r_c,r_{(c+1)},---,r_{(4m+1)c})$

$(1+x)a(x)=(A,A,A,---,A)$, where there are $(4m+1)$A's and $A=(a_1,a_1+a_2,---,a_{(c-2)}+a_{(c-1)},a_{(c-1)})$.

The decoding of the multiple solid-burst-error-correcting cyclic codes can be proceeded as:

Since any one of the set $\{r_i, r_{(c+i)}, r_{(2c+i)}, ---, r_{(4mc+i)}\}$, $1 \leq i \leq c$ contains at most 2m multiple random errors. There are at least (2m+1) correct digits in the set. Since $2m+1 > [\frac{4m+1}{2}]$, $a_1, a_i+a_{(i-1)}$, $2 \leq i \leq c-1$ and $a_{(c-1)}$ can be decoded by the majority decisions as follows:

$$a_1 = \text{maj.}\{r_1, r_{(c+1)}, r_{(2c+1)}, ---, r_{(4mc+1)}\}$$

$$a_i+a_{(i-1)} = \text{maj.}\{r_i, r_{(c+i)}, ---, r_{(4mc+i)}\}, \quad 2 \leq i \leq c-1.$$

$$a_{(c-1)} = \text{maj.}\{r_c, r_{2c}, r_{3c}, ---, r_{(4m+1)c}\}.$$

As a result of lemma 1, all the information digits can be computed.

Example 1 will make the decoding algorithm of a multiple solid-burst-error-correcting cyclic codes clear.

(Example 1)

$g(x) = \sum_{i=0}^{12} x^{5i}$ generates a cyclic code of length n=65, which is capable of correcting triple solid-burst-errors of arbitrary length. The last digit of the information digits is always assumed to be zero. $I = (a_1, a_2, a_3, a_4)$ is the set of the actual useful information digits.

Then $\{a_1, a_1+a_2, a_2+a_3, a_3+a_4$, and $a_4\}$ can be determined as follows:

$$a_1 = \text{maj.}(r_1, r_6, r_{11}, --- , r_{61})$$

$$a_1+a_2 = \text{maj.}(r_2, r_7, r_{12}, --- , r_{62})$$

$$a_2+a_3 = \text{maj.}(r_3, r_8, r_{13}, --- , r_{63})$$

$$a_3+a_4 = \text{maj.}(r_4, r_9, r_{14}, -- , r_{64})$$

$$a_4 = \text{maj.}(r_5, r_{10}, r_{15}, --- , r_{65})$$

As a result of lemma 1, all the information digits $\{a_1, a_2, a_3, a_4\}$ can be computed.

If the decoding algorithm is modified, then the number of check digits can be significantly reduced for the single and double solid-burst-error-correcting codes. The single and double solid-burst-error-correcting cyclic codes presented are similar to the result of Shiva and Sheng (1969). However, the decoding algorithm is completely different. Besides, the error correction capability is better than Shiva and Sheng's result. The following will present single and double solid-burst-error-correcting cyclic codes.

(B) Single Solid-Burst-Error Correcting Cyclic Codes

(Theorem 2) $g(x)=1+x^c+x^{2c}$ generates a cyclic code of length n=3c which is capable of correcting all the single solid-burst-errors of arbitrary length provided that the actual useful information digits is (c-1).

The received code polynomial $r(x)$ can be expresses as: $r(x)=a(x)+B_1(x)$, where $B_1(x)$ is a single solid burst error pattern of arbitrary length. $(1+x)r(x)=(1+x)a(x)+x^{j_1}+x^{j_2}$, where j_1 and j_2 are two non-negative integers. $(1+x)r(x)$ will be the actual received code polynomial used in decoding. The following decoding algorithm is obtained as a result of these two random errors x^{j_1} and x^{j_2}. $(1+x)r(x)$ can be expressed as $(1+x)r(x)=(r_1,r_2,r_3,---,r_c,r_{(c+1)},---,r_{2c},r_{(2c+1)},---,r_{3c})$. The decoding algorithm is:

(Step 1) If there are at least $3(c-1)$ digits of $(1+x)r(x)$ which satisfy the condition $r_i=r_{(c+i)}=r_{(2c+i)}$, $1 \leq i \leq c$, then at least $(c-1)$ digits of the set $\{a_1,a_1+a_2,a_2+a_3,a_3+a_4,---,a_{(c-2)}+a_{(c-1)},a_{(c-1)}\}$ can be decoded as:

$a_1=r_1, a_i+a_{(i-1)}=r_i$, $2 \leq i \leq (c-1)$ and $a_{(c-1)}=r_c$. As a result of lemma 1, all the information digits can be computed.

(Step 2) If step 1 fails, then all the digits in the set $\{a_1,a_1+a_2,a_2+a_3,---,a_{(c-2)}+a_{(c-1)},a_{(c-1)}\}$ can be decoded by majority decisions as:

$a_1=\text{maj.}(r_1,r_{(c+1)},r_{(2c+1)})$

$a_i+a_{(i-1)}=\text{maj.}\ (r_i,\ r_{(c+i)},r_{(2c+i)})$, $2 \leq i \leq (c-1)$.

$a_{(c-1)}=\text{maj.}\ (r_c,r_{2c},r_{3c})$.

As a result of lemma 1, all the information digits can be computed.

The following will show that the single solid-burst-error-correcting cyclic codes can indeed be decoded by the two steps which are described as above.

(CASE 1) If the single solid-burst-error pattern spans the whole code length, then $r(x)=a(x)+\sum_{i=0}^{(n-1)} x^i$. Therefore $(1+x)r(x)=(1+x)a(x)$ (mod x^n+1). Therefore all the 3c digits of $(1+x)r(x)$ satisfy the condition $r_i=r_{(c+i)}=r_{(2c+i)}$, $1 \leq i \leq c$. Therefore the received code polynomial can be decoded by step 1.

(CASE 2) If $B_1(x) = x^j\ (\sum_{i=0}^{(\ell-1)} x^i)$, then $(1+x)r(x)=(1+x)a(x)+x^j+x^{j+\ell}$, where ℓ is the length of the solid-burst-error pattern. There are two possibilities which have to be considered: (1°) If the length of the solid-burst-error pattern ℓ is c or 2c, then $x^i+x^{i+\ell}=x^i+x^{i+c}$ or x^i+x^{i+2c}. For this possibility, the set $\{r_{(i+1)},r_{(c+i+1)},r_{(2c+i+1)}\}$ contains the two random errors $x^i+x^{(i+\ell)}$. Therefore there are exactly $3(c-1)$ digits of $(1+x)r(x)$ which satisfy the condition $r_j=r_{(j+c)}=r_{(j+2c)}$, $1 \le j \le c$. Therefore the received code polynomial $(1+x)r(x)$ can be decoded by step 1.

(2°) If $\ell \neq c$ and $\ell \neq 2c$, then any set $\{r_j, r(j+c), r(j+2c)\}$, $1 \le j \le c$ contains at most a single error, therefore there are at least two correct digits in each set $\{r_j, r(j+c), r(j+2c)\}$. Therefore all the digits in the set $\{a_1, a_1+a_2, a_2+a_3, ---, a_{(c-2)}+a_{(c-1)}, a_{(c-1)}\}$ can be decoded by majority decisions.

(C) Double Solid-Burst-Error Correcting Cyclic Codes

(Theorem 3) $g(x)=1+x^c+x^{2c}+x^{3c}+x^{4c}$ generates a cyclic code of length $5c$ which is capable of correcting double solid-burst-errors of arbitrary length provided that the actual useful information digits is $(c-1)$ digits.

The received code polynomial $r(x)$ can be expressed as

$r(x)=a(x)+B_1(x)+B_2(x)$, where $B_1(x)$ and $B_2(x)$ are solid burst error patterns of arbitrary length. $(1+x)r(x)=(1+x)a(x)+x^{i_1}+x^{i_2}+x^{i_3}+x^{i_4}$, where i_1, i_2, i_3, i_4 are four arbitrary positive integers. $(1+x)r(x)$ contains at most four random errors.

Let $(1+x)r(x)=(r_1, r_2, r_3, ---, r_c r_{(c+1)}, ---, r_{4c}, ---, r_{5c})$. If $(1+x)B(x)=0$, then there are five correct decision for determining each digit of the set $\{a_1, a_i+a_{(i-1)}, 2 \le i \le c-1, a_{(c-1)}\}$.

They are:

$a_1=r_1=r_{(c+1)}=r_{(2c+1)}=r_{(3c+1)}=r_{(4c+1)}$, and

$a_i+a_{(i-1)}=r_i=r_{(c+i)}=r_{(2c+i)}=r_{(3c+i)}=r_{(4c+i)}$, $2 \le i \le (c-1)$, and

$a_{(c-1)}=r_c=r_{2c}=r_{3c}=r_{4c}=r_{5c}$.

If $(1+x)B(x) \neq 0$, then $(1+x)r(x)$ can be decoded as:

(Step 1) If one of the subset $\{r_i, r_{(i+c)}, r_{(i+2c)}, r_{(i+3c)}, r_{(i+4c)}\}$, where $1 \le i \le c$ contains all the four random errors $\{x^{i_1}, x^{i_2}, x^{i_3}, x^{i_4}\}$, then there are at least $(c-1)$ subsets $\{r_i, r_{(i+c)}, r_{(i+2c)}, r_{(i+3c)}, r_{(i+4c)}\}$, $1 \le i \le c$ which are free from any errors. For these $(c-1)$ subsets which are free from errors, all the five digits in the subset must be equal i.e. $r_i=r_{(i+c)}=r_{(i+x)}=r_{(i+3c)}=r_{(i+4c)}$. These $(c-1)$ subsets can be used to decode the $(c-1)$ digits of the set $\{a_1, a_i+a_{(i-1)}, 2 \le i \le c-1, a_{(c-1)}\}$ as follows:

$a_1=r_1$, $a_i+a_{(i-1)}=r_i$, $2 \le i \le c-1$, and $a_{(c-1)}=r_c$.

As a result of lemma 1, all the information digits can be computed.

If step 1 fails, then go to step 2.

(Step 2) If one of the subset $\{r_i, r_{(i+c)}, r_{(i+2c)}, r_{(i+3c)}, r_{(i+4c)}\}$ $1 \leq i \leq c$ contains 3 random errors, then in the remaining (c-1) subsets, there must be at least 4 elements in each subset which are equal to each other. These (c-1) subsets which have four elements equal to each other can be used to decode the (c-1) digits in the set $\{a_1, a_i + a_{(i-1)}, 2 \leq i \leq c-1, a_{(c-1)}\}$.

Define $f(x_1,x_2,x_3,x_4,x_5)$ to be a function such that $f(x_1,x_2,x_3,x_4,x_5)$ $=1$ if there are at least four out of the five x_i's which have values 1. $f(x_1,x_2,x_3,x_4,x_5)=0$ if there are at least four out of the five x_i's which have values 0.

The (c-1) digits of the set $\{a_1, a_i + a_{(i-1)}, 2 \leq i \leq (c-1)$ and $a_{(c-1)}\}$ are decoded as:

$$a_1 = f(r_1, r_{(c+1)}, r_{(2c+1)}, r_{(3c+1)}, r_{(4c+1)})$$

$$a_i + a_{(i-1)} = f(r_i, r_{(c+i)}, r_{(2c+i)}, r_{(3c+i)}, r_{(4c+i)}), \; 2 \leq i \leq (c-1)$$

$$a_{(c-1)} = f(r_c, r_{2c}, r_{3c}, r_{4c}, r_{5c}).$$

Therefore at least (c-1) digits of the set $\{a_1, a_i + a_{(i-1)}, 2 \leq i \leq (c-1), a_{(c-1)}\}$ can be obtained. As a result of lemma 1, all the information digits can be computed.

(Step 3) If both steps 1 and 2 fail, then none of the subsets $\{r_i, r_{(i+c)}, r_{(i+2c)}, r_{(i+3c)}, r_{(i+4c)}\}$, $1 \leq i \leq c$ should contain more than double errors. Therefore in each subset $\{r_i, r_{(i+c)}, r_{(i+2c)}, r_{(i+3c)}, r_{(i+4c)}\}$, there are at least three correct digits. Therefore each digit in the set $\{a_1, a_i + a_{(i-1)}, 2 \leq i \leq (c-1)$ and $a_{(c-1)}\}$ can be decoded as:

$$a_1 = \text{maj.}\ \{r_1, r_{(c+1)}, r_{(2c+1)}, r_{(3c+1)}, r_{(4c+1)}\}$$

$$a_i + a_{(i-1)} = \text{maj.}\ \{r_i, r_{(c+i)}, r_{(2c+i)}, r_{(3c+i)}, r_{(4c+i)}\}, \; 2 \leq i \leq c-1$$

$$a_{(c-1)} = \text{maj.}\ \{r_c, r_{2c}, r_{3c}, r_{4c}, r_{5c}\}$$

As a result of lemma 1, all the information digits can be computed.

REFERENCES

R. T. Chien, K. M. Tzeng (1969), "Cyclic Codes for Solid Burst Error Correction", Proceedings of the Third Annual Princeton Conference on Information Sciences and Systems.

H. T. Hsu (1968), "A Class of Solid Burst Error Correcting Cyclic Codes", Proceedings of the National Electronics Conference, pp. 402-404, Vol. 24, 1968.

A. G. Schillinger (1964), " A Class of Solid Burst Error Correcting Codes", Polytechnic Institute of Brooklyn, N.Y., Research Dept. PIBMRI-1223-64.

S.G.S. Shiva, C.L. Sheng (1969), "Multiple Solid Burst-Error-Correcting Binary Codes", IEEE Trans. Information Theory, Vol. IT-15, No. 1, pp. 188-189.

ON REDUCING LINEAR DISCRETE SYSTEM STATE ESTIMATION COMPUTER REQUIREMENTS

ROLAND PRIEMER
University of Illinois
Chicago, Illinois

ABSTRACT

A procedure is presented for reducing the computing time and storage requirements when estimating the state of a class of linear discrete systems.

INTRODUCTION

Consider the problem of estimating the state of a linear discrete system described by

$$x(k + 1) = \phi_0 x(k) + B_0 u(k) \tag{1}$$

where the state x is an n vector, ϕ_0 and B_0 are known n x n and n x 1 matrices respectively, and u(k), a scalar, is the random system distrubance. The initial state is a random vector with zero mean and covariance V_x. Usually, the measurement process, which is assumed to be linear, is given and is described by

$$y(k) = D_0 x(k) + w(k) \tag{2}$$

where the measurement y is an m vector, $m \leq n$, and w(k) is the measurement noise. The random distrubances u(k) and w(k) are independent and both are zero mean white sequences which are independent of the initial state. Since D_0, the rows of which are linearly independent, is given, the results of Kalman [1] and Rauch et al.[2] can be used to solve the filtering and smoothing problems respectively. These are linear estimators, which yield the estimate $\hat{x}(k/\ell)$ of the state x(k) given the measurements $Y(\ell) = \{y(i), i = 0, \ldots, \ell\}$, such that the mean square error is minimized.

Now suppose that a part of the implementation problem is to reduce the computing time and storage requirements for such estimators. Computing restrictions such as these may naturally arise when estimates are required in real time or the estimator is to be realized via hardware. In [3]-[6] various aspects of the estimation problem have been considered, which resulted in techniques that reduce computing requirements. In the present paper, a procedure is developed which reduces a class of linear $n^{\underline{th}}$ order systems to $m^{\underline{th}}$ order systems with time delays, and therefore allows the application of optimal linear estimators developed by Priemer and Vacroux [7], [8], which require less computing time and memory than optimal linear estimators applied to the $n^{\underline{th}}$ order system. In this procedure D_0 is considered to be a design parameter rather than as a given part of the estimation problem.

THE MODIFIED ESTIMATION PROCEDURE

Suppose a nonsingular transformation T exists, such that: v(k) = Tx(k). Therefore, (1) and (2) become

$$v(k + 1) = \phi v(k) + Bu(k) \tag{3}$$

$$y(k) = Dv(k) + w(k) \tag{4}$$

Athans [9] has shown that the optimum linear estimates of the states of alternate state-space representations of continuous linear systems, which are related by a nonsingular linear transformation, are also related by that linear transformation. Similarly, it can be shown for linear discrete systems that if T is nonsingular, then $\hat{x}(k/\ell) = T^{-1}\hat{v}(k/\ell)$. In obtaining $\hat{v}(k/\ell)$, the following will be required, where for convenience m = 1,

$$\phi = T\,\phi_0\,T^{-1} = \begin{bmatrix} a_o & a_1 & \dots & a_{n-1} \\ 1 & & & 0 \\ & \ddots & & \\ 0 & & 1 & \end{bmatrix}, \qquad B = T\,B_0 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (5), (6)$$

$$D = D_0\,T^{-1} = [\,d \; 0 \; \dots \; 0\,]. \qquad (7)$$

If T_n denotes the last row of T, then by a procedure somewhat similar to that of Ramaswami and Ramar [10], it can be shown that T_n must satisfy

$$T_n C = T_n [B_0 \;\; \phi_0 B_0 \; \dots \; \phi_0^{n-1} B_0] = [0 \dots 0 \; 1]. \qquad (8)$$

If C^{-1} exists, then the remaining rows of T are generated by

$$T_{n-i} = T_{n-(i-1)}\phi_0 \qquad (9)$$

for $i = 1, \dots, n - 1$. Once T has been determined, then (5) is used to determine the a_i, and by (7): $D_0 = d\,T_1$. Note that some flexibility in achieving a specified mean square estimation error is allowed, because d is still an unconstrained parameter.

The structure of ϕ, B, and D enables (3) and (4) to be written in another way. Let $z(k)$ be defined by: $v'(k) = [z(k) \; z(k - 1) \; \dots \; z(k - (n-1))]$, and therefore the process $v(k)$ is completely described by

$$z(k + 1) = \sum_{j=0}^{n-1} a_j z(k - j) + u(k) \qquad (10)$$

$$y(k) = d\,x(k) + w(k) \qquad (11)$$

The problem of estimating the state $x(k)$ given $Y(\ell)$ has now been transformed to obtaining the estimates $\hat{z}(k - j/\ell)$ of the states $z(k - j)$, $j = 0, \dots, n - 1$ given $Y(\ell)$ of the scalar process described by (10). It is easily verified by example, for those linear discrete systems for which C^{-1} exists, that the estimators obtained in [7] and [8], when using the parameters of (10) and (11), significantly reduce the computing requirements of obtaining $\hat{x}(k/\ell)$. The above technique can be extended to measurement processes of order other than one for those m such that n is an integer multiple of m. If this is not the case, then augmenting the model of the system by r states such that $n + r$ is an integer multiple of m merits consideration.

REFERENCES

1. R.E. Kalman, "A new approach to linear filtering and prediction problems", Trans. ASME, J. Basic Engrg., ser. D, Vol. 82, pp. 35-45, 1960.
2. H.E. Rauch, F. Tung, and C.T. Striebel, "Maximum likelihood estimates of linear dynamic systems", AIAA J., vol. 3, no. 8, pp. 1445-1450, 1965.
3. C.F. Price, "An analysis of the divergence problem in the Kalman filter", IEEE Trans. A-C, pp. 699-702, Dec., 1968.
4. C.H. Wells, "An approximation of the Kalman filter equations", IEEE Trans. A-C, pp. 445, Aug., 1968.
5. K.C. Brammer, "Lower order optimal linear filtering of nonstationary random sequences", IEEE Trans. A-C, pp. 198-199, April, 1968.
6. K.W. Simon and A.R. Stubberud, "Reduced order Kalman filter", Int. J. Control, pp. 501-509, Nov., 1969.
7. R. Priemer and A.G. Vacroux, "Estimation in linear discrete systems with multiple time delays", IEEE Trans. A-C, pp.384-387, Aug., 1969.
8. R. Priemer and A.G. Vacroux, "On smoothing in linear discrete systems with time delays", Int. J. Control, to be published, 1970.
9. M. Athans, "The relationship of alternate state space representations in linear filtering problems", IEEE Trans. A-C, pp. 775-776, Dec., 1967.
10. B. Ramaswami and K. Ramar, "On the transformation of time variable systems to phase variable canonical form", IEEE Trans. A-C, pp. 417-419, Aug., 1969.

KALMAN FILTER FORMULATION OF AN ALTITUDE RATE ESTIMATOR FOR AIRCRAFT FLYING AUTOMATIC TERRAIN FOLLOWING

GAYLORD W. CARLOCK
General Dynamics
Fort Worth, Texas

ABSTRACT

The formulation of an altitude rate estimator for aircraft flying automatic terrain following is presented. The estimator results from an application of the extended Kalman filter to estimate the radar altitude rate from measurements made of the radar altitude. A nonlinear dynamic model is formulated for the system; the first-order nonlinear approximate conditional-mean filter equations are manipulated into a suitable format; and the a priori statistics, initial conditions, and check-out procedure are discussed.

INTRODUCTION

The control system for an aircraft flying in an automatic terrain following mode typically utilizes radar altimeter rate information in the generation of flight control commands. The rate information is commonly obtained from differentiation of radar altimeter measurements.

This paper discusses a Kalman filter formulation of an altitude rate estimator which utilizes measurements of radar altitude as well as a priori statistics to obtain the rate estimate. In a degenerate case where the optimal gains take on specific values, the formulation is seen to reduce to a simple differentiating circuit followed by a low-pass filter.

The formulation of the estimator presented here represents a "first pass" design and its applicability has not been substantiated.

MODEL EQUATIONS AND PHILOSOPHY

A critical step in the derivation of the estimator is the construction of system and observation models which adequately describe the "real world" process. Nonlinear model equations are usually state variable time domain expressions of the form [1,2,3,4].

$$\dot{x}(t) = f\,[\,x(t),\, t\,] + G\,[\,x(t),\, t\,]\, w(t) \tag{1a}$$

$$z(t) = h\,[\,x(t),\, t\,] + v(t) \tag{1b}$$

where

$x(t)$ is the $(n \times 1)$ system state vector

$z(t)$ is the $(m \times 1)$ observation vector

f and h are memoryless, possibly nonlinear functions of $x(t)$ and t

G is a $(n \times r)$ memoryless, possibly nonlinear function of $x(t)$ and t

$w(t)$ is the $(r \times 1)$ system noise vector

$v(t)$ is the $(m \times 1)$ observation noise vector and w(t) and v(t) are zero-mean, white, and gaussian such that

$$\mathrm{Cov}\,[w(t),\, w(\tau)] = \psi_w(t)\ \delta_D(t-\tau)$$

$$\mathrm{Cov}\,[v(t),\, v(\tau)] = \psi_v(t)\ \delta_D(t-\tau)$$

and $\mathrm{Cov}\,[v(t),\, w(\tau)] = 0$

It is theoretically desirable to define the dynamic system model with as many state variables as the number of variables accounting for the "activity" of the system. However, computer limitations and implementation details usually restrict the number of variables defined [5,6]. The model must be sophisticated enough to account for all the essential uncertainties while simple enough to meet implementation constraints.

In the mathematical formulation that follows, the detailed state equations of the automatic flight control/aircraft/aerodynamic system will not be utilized. Instead, a simpler approach will be taken which exploits the equations which define the basic nature of automatic terrain following flight. In this development, a distinction between aircraft pressure altitude and aircraft radar altitude is required. Aircraft pressure altitude will be used to mean the aircraft altitude referenced to sea level. Aircraft radar altitude will be used to mean the aircraft altitude referenced to the terrain surface directly below the aircraft

FORMULATION OF THE SYSTEM MODEL

The function of automatic terrain following (ATF) flight is to provide aircraft low-level penetration capability. The control system commands the aircraft to fly at a designated set clearance, H_o, above the terrain contingent upon satisfying certain state variable constraint equations. During ATF flight, the terrain directly underneath the aircraft changes as a function of time.

Consider the situation depicted in Figure 1 where the aircraft is at its designated set clearance, H_o, above the terrain at time t_1. If the aircraft is to remain at the set clearance at time $t_1 + \Delta t$, then the vertical displacement of the aircraft must exactly equal the change in the terrain altitude during the interval $(t_1,\ t_1 + \Delta t)$. Thus the ideal ATF equation for $\Delta t \rightarrow 0$ can be written

$$U(t)\ \sin \gamma_I(t)\,\Delta t = \dot{h}_t(t)\,\Delta t \tag{2}$$

where

U = aircraft speed

γ_I = ideal flight vector angle

$\dot{h}_t$ = time rate of change of terrain altitude

Solving Eq. (2) for $\gamma_I(t)$, the ideal flight vector angle can be expressed

$$\gamma_I(t) = \sin^{-1}\left[\frac{\dot{h}_t(t)}{U(t)}\right] \tag{3}$$

Actual implementation of the ATF control system and aircraft/system operational constraints result in the aircraft flying with an actual flight vector angle, $\gamma_A(t)$, which may differ from the ideal angle, i e.,

$$\gamma_A(t) = \gamma_I(t) + \gamma_\epsilon(t) \tag{4}$$

where

γ_A = actual flight vector angle

γ_I = ideal flight vector angle

γ_ϵ = error

Thus, the time rate of change of the aircraft in the vertical direction can be written

$$\dot{h}_a(t) = U(t) \sin \gamma_A(t) \tag{5}$$

where

$h_a(t)$ = aircraft pressure altitude

Substituting Eq. (4) into Eq. (5), the following expression is obtained.

$$\dot{h}_a(t) = U(t) \sin \left[\gamma_I(t) + \gamma_\epsilon(t)\right] \tag{6}$$

Since aircraft pressure altitude is referenced to sea level and aircraft radar altitude is referenced to the terrain surface, we can write

$$h_a(t) = h_R(t) + h_t(t) \tag{7}$$

where

h_a = aircraft pressure altitude

h_R = aircraft radar altitude

h_t = terrain altitude

Differentiating term-by-term and solving for $\dot{h}_R(t)$

$$\dot{h}_R(t) = \dot{h}_a(t) - \dot{h}_t(t) \tag{8}$$

Substituting Eq. (6) into Eq. (8)

$$\dot{h}_R(t) = U(t) \sin\left[\gamma_I(t) + \gamma_\epsilon(t)\right] - \dot{h}_t(t) \tag{9}$$

Substituting Eq. (3) into Eq. (9)

$$\dot{h}_R(t) = U(t) \sin\left[\sin^{-1}\left[\frac{\dot{h}_t(t)}{U(t)}\right] + \gamma_\epsilon(t)\right] - \dot{h}_t(t) \tag{10}$$

Using the identities $\sin(x+y) = \sin x \cos y + \cos x \sin y$ and $\cos(\sin^{-1}x) = (1 - x^2)^{\frac{1}{2}}$, Eq. (10) can be rewritten as

$$\dot{h}_R(t) = \dot{h}_t(t) \cos \gamma_\epsilon(t) + U(t) \left[1 - \left[\frac{\dot{h}_t(t)}{U(t)} \right]^2 \right]^{\frac{1}{2}} \sin \gamma_\epsilon(t) - \dot{h}_t(t) \qquad (11)$$

Making the following assumptions

$$\begin{aligned} &(1) \quad \mathrm{Var}[\gamma_\epsilon(t)] < 10 \text{ (degrees)}^2 \\ &(2) \quad \mathrm{Var}[\dot{h}_t(t)] < 0.2\, U^2(t) \quad (\text{ft/sec})^2 \end{aligned} \qquad (12)$$

we can write

$$\cos \gamma_\epsilon(t) \cong 1$$

$$\sin \gamma_\epsilon(t) \cong \gamma_\epsilon(t)$$

$$\left[1 - \left[\frac{\dot{h}_t(t)}{U(t)} \right]^2 \right]^{\frac{1}{2}} \cong 1 - \frac{1}{2} \left[\frac{\dot{h}_t(t)}{U(t)} \right]^2$$

and by substitution into Eq. (11) we obtain

$$\dot{h}_R(t) = U(t) \left[1 - 0.5 \left[\frac{\dot{h}_t(t)}{U(t)} \right]^2 \right] \gamma_\epsilon(t)$$

$$= U(t)\, \gamma_\epsilon(t) - \frac{0.5}{U(t)} \left[\dot{h}_t(t) \right]^2 \gamma_\epsilon(t) \qquad (13)$$

Uncertainty enters the equation for radar altitude rate, $\dot{h}_R(t)$, through the quantities $\dot{h}_t(t)$ and $\gamma_\epsilon(t)$. These quantities, respectively, are the rate of change of the terrain altitude and the difference between the actual aircraft flight vector angle and the ideal ATF flight vector angle. Intuition, as well as observed data, indicates that these quantities may be mathematically modeled as time-correlated random processes. Here they are assumed to be correlated exponentially with themselves but uncorrelated with each other so that

$$\mathrm{Cov}\left[\dot{h}_t(t), \dot{h}_t(\tau) \right] = A_1 \exp[-a_1 \, |t-\tau|]$$

$$\mathrm{Cov}\left[\gamma_\epsilon(t), \gamma_\epsilon(\tau) \right] = A_2 \exp[-a_2 \, |t-\tau|]$$

and $\mathrm{Cov}\left[\dot{h}_t(t), \gamma_\epsilon(\tau) \right] = 0$

It can be easily shown that the above covariance expressions result for $\dot{h}_t(t)$ and $\gamma_\epsilon(t)$ when they are modeled as the output of first order linear systems driven by independent white noise sources:

$$\ddot{h}_t(t) = -a_1 \, \dot{h}_t(t) + w_1(t) \qquad (14a)$$

$$\dot{\gamma}_\epsilon(t) = -a_2 \, \gamma_\epsilon(t) + w_2(t) \qquad (14b)$$

where

$w_1(t)$ and $w_2(t)$ are zero-mean, white, gaussian noise sources with

$$\mathrm{Cov}\,[\,w_1(t),\ w_1(\tau)\,] = 2a_1\ A_1\ \delta(t-\tau)$$

$$\mathrm{Cov}\,[\,w_2(t),\ w_2(\tau)\,] = 2a_2\ A_2\ \delta(t-\tau)$$

$$\mathrm{Cov}\,[\,w_1(t),\ w_2(\tau)\,] = 0$$

The dynamic system model can now be formulated into the format of Equation (1a) by utilizing the foregoing equations.

Letting

$$x_1 = \dot{h}_t(t)$$

$$x_2 = \gamma_\epsilon(t)$$

$$x_3 = h_R(t)$$

Equations (13), (14a), and (14b) can be rewritten as

$$\dot{x}_1 = -\ a_1\ x_1 + w_1(t)$$

$$\dot{x}_2 = -\ a_2\ x_2 + w_2(t)$$

$$\dot{x}_3 = U(t)\ x_2 - \frac{0.5}{U(t)}\ x_1{}^2\ x_2$$

and finally in matrix form as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -\ a_1\ x_1 \\ -\ a_2\ x_2 \\ U(t)\ x_2 - \frac{0.5}{U(t)}\ x_1{}^2\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \tag{15}$$

where from comparison with Equation (1a) it is seen that $n = 3$, $r = 2$, and

$$f[x(t),\ t] = \begin{bmatrix} -\ a_1\ x_1 \\ -\ a_2\ x_2 \\ U(t)\ x_2 - \frac{0.5}{U(t)}\ x_1{}^2\ x_2 \end{bmatrix}$$

$$G[x(t),\ t] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$\mathrm{Cov}\left[\, w(t),\, w(\tau) \,\right] = \psi_w \, \delta_D(t-\tau) = \begin{bmatrix} 2a_1 \, A_1 \, \delta_D(t-\tau) & 0 \\ 0 & 2a_2 \, A_2 \, \delta_D(t-\tau) \end{bmatrix}$$

OBSERVATION MODEL FORMULATION

The one quantity that is assumed to be measured in this formulation is the radar altitude, $h_R(t)$. It is assumed that this measured value is contaminated with zero-mean, white, gaussian noise. Thus, the observation model can be written

$$z(t) = x_3 + v(t) \tag{16}$$

where

z = measured value of radar altitude

x_3 = true value of radar altitude

v = zero-mean, white, gaussian noise

From comparison with Equation (1b) it is seen that m = 1 and

$$h\left[\, x(t),\, t \,\right] = x_3$$

$$\mathrm{Cov}\left[\, v(t),\, v(\tau) \,\right] = \psi_v \, \delta_D(t-\tau)$$

MANIPULATION OF THE NONLINEAR ESTIMATION EQUATIONS

The first-order nonlinear approximate conditional-mean filter algorithm may be written [4]

$$\dot{\hat{x}}(t) = f[\hat{x},t] + V_{\tilde{x}}(t) \frac{\partial}{\partial \hat{x}} h^T[\hat{x},t] \, \psi_v^{-1} \left[z(t) - h[\hat{x},t] \right] \tag{17}$$

and the error-variance algorithm may be written

$$\begin{aligned} \dot{V}_{\tilde{x}}(t) = {} & \frac{\partial}{\partial \hat{x}} f[\hat{x},t] \, V_{\tilde{x}} + V_{\tilde{x}} \frac{\partial}{\partial \hat{x}} f^T[\hat{x},t] \\ & + G[\hat{x},t] \, \psi_w \, G^T[\hat{x},t] \\ & - V_{\tilde{x}} \frac{\partial}{\partial \hat{x}} h^T[\hat{x},t] \, \psi_v^{-1} \frac{\partial}{\partial \hat{x}} h[\hat{x},t] \, V_{\tilde{x}} \end{aligned} \tag{18}$$

Utilizing the previously derived expressions for $f[x,t]$, $G[x,t]$, $h[x,t]$, ψ_w, and ψ_v and carrying out the indicated operations, Equation (17) for this formulation becomes

$$\begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \dot{\hat{x}}_3 \end{bmatrix} = \begin{bmatrix} -\, a_1 \, \hat{x}_1 \\ -\, a_2 \, \hat{x}_2 \\ U(t) \, \hat{x}_2 - \frac{0.5}{U(t)} \, \hat{x}_1^2 \, \hat{x}_2 \end{bmatrix} + \begin{bmatrix} V_{13}(t) \\ V_{23}(t) \\ V_{33}(t) \end{bmatrix} \left[\frac{z(t) - \hat{x}_3}{\psi_v} \right] \tag{19}$$

where from Equation (18) we can write

$$\dot{V}_{11} = -2\, a_1\, V_{11} + 2\, a_1\, A_1 - \psi_v^{-1}\, V_{13}^{\;2}$$

$$\dot{V}_{12} = -(a_1 + a_2)\, V_{12} - \psi_v^{-1}\, V_{13}\, V_{23}$$

$$\dot{V}_{13} = -\, a_1\, V_{13} - \frac{\hat{x}_1\, \hat{x}_2}{U(t)}\, V_{11} + \left[U(t) - \frac{\hat{x}_1^{\,2}}{2U(t)}\right] V_{12} - \psi_v^{-1}\, V_{13}\, V_{33}$$

$$\dot{V}_{22} = -2\, a_2\, V_{22} + 2\, a_2\, A_2 - \psi_v^{-1}\, V_{23}^{\;2}$$

$$\dot{V}_{23} = -\, a_2\, V_{23} - \frac{\hat{x}_1\, \hat{x}_2}{U(t)}\, V_{12} + \left[U(t) - \frac{\hat{x}_1^{\,2}}{2U(t)}\right] V_{22} - \psi_v^{-1}\, V_{23}\, V_{33}$$

$$\dot{V}_{33} = -\,\frac{2\, \hat{x}_1\, \hat{x}_2}{U(t)}\, V_{13} + 2\left[U(t) - \frac{\hat{x}_1^{\,2}}{2U(t)}\right] V_{23} - \psi_v^{-1}\, V_{33}^{\;2}$$

which are elements of the symmetric matrix, $\dot{V}_{\tilde{x}}$.

Using Equation (19), the block diagram shown in Figure 2 can be drawn. The optimal gains $K_1(t)$, $K_2(t)$, and $K_3(t)$ are time varying gains which are computed continuously from Equations (17) and (18). Their computation constitutes the major complication of the formulated estimator. Allowing a sub-optimal solution with the gains set at some constant value would greatly reduce the implementation complexity with, perhaps, only a small degradation in performance. Note that when $K_1(t)$ and $K_2(t)$ are equal to zero, the estimator reduces to a simple differentiating circuit followed by a low-pass filter as shown in Figure 3.

INITIALIZATION AND CHECKOUT

Utilization of the formulated estimator requires an initialization scheme for the estimator and the establishment of variance parameters for the noise sources. A reasonable scheme would be to initialize the state variables over level ground at the beginning of the ATF segment of flight with $x_1(0) = x_2(0) = 0$ and $x_3(0) = H_o$. Initial variances would be established after terrain and system analysis had been performed.

Ensemble characteristics of the estimator may be investigated on a digital computer utilizing the previously derived equations; however, the real test of the estimator requires the use of a complex hybrid computer to simulate the entire aircraft/ATF system together with the estimation algorithm. The simulation should include an operational spectrum composed of a number of aircraft configurations, speeds, and different terrains. This would determine the adequacy of the derived system model as well as the estimation equations.

CONCLUSIONS

A Kalman filter altitude rate estimator was formulated using a rather simple dynamic system model. The adequacy of the model and the initial conditions and variance parameters will need to be determined by a complex hybrid computer simulation of the entire system. It was shown that under certain conditions, the estimator degenerates into a differentiator followed by a low-pass filter.

ACKNOWLEDGMENT

The author is grateful to J. J. Errante for enabling the preparation of this paper and to Janis Thomas for providing typing service.

REFERENCES

1. Jazwinski, A. H., "Filtering for Nonlinear Dynamical Systems", IEEE Transactions on Automatic Control (Correspondence), Vol AC-11, No. 4, October, 1966.

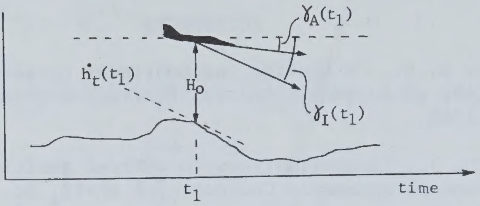
2. Kushner, H. J., "Approximations to Optimal Nonlinear Filters", IEEE Transactions on Automatic Control, Vol AC-12, No. 5, October, 1967.

3. Synder, D. L., State Variable Approach to Continuous Estimation with Applications to Analog Communication Theory, MIT Monograph #51, 1969.

4. Sage, A. P., and Melsa, J. L., Estimation Theory with Applications to Communications and Control, McGraw-Hill Book Company, New York, 1970.

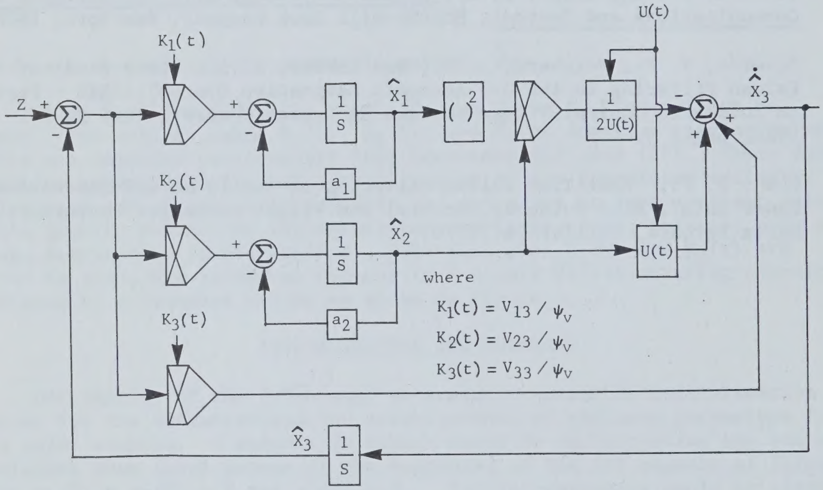
5. Schmidt, S. F., Weinberg, J. D., and Lukesh, J. S., "Case Study of Kalman Filtering in the C-5 Aircraft Navigation System", IEEE - Group on Automatic Control - Case Studies in System Control, June 24-25, 1968, pages 59-109.

6. Chen, P. P., "Real-Time Kalman Filtering of Apollo LM/AGS Rendezvous Radar Data", AIAA Guidance, Control and Flight Mechanics Conference, Santa Barbara, California, 1970.



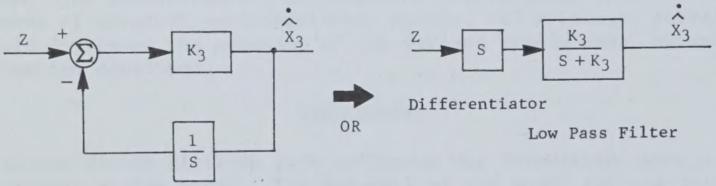
$\gamma_A(t_1)$
$\dot{h}_t(t_1)$
H_o
$\gamma_I(t_1)$
t_1
time

Figure 1 GEOMETRY FOR ATF FLIGHT



U(t)
$K_1(t)$
Z
+
-
$\hat{X}_1$
$\frac{1}{S}$
$(\ )^2$
$\frac{1}{2U(t)}$
$\hat{\dot{X}}_3$
a_1
$K_2(t)$
$\hat{X}_2$
U(t)
a_2
$K_3(t)$
where
$K_1(t) = V_{13} / \psi_v$
$K_2(t) = V_{23} / \psi_v$
$K_3(t) = V_{33} / \psi_v$
$\hat{X}_3$
$\frac{1}{S}$

Figure 2 KALMAN ALTITUDE RATE ESTIMATOR



Z
+
-
K_3
$\hat{\dot{X}}_3$
$\frac{1}{S}$
OR
Z
S
$\frac{K_3}{S + K_3}$
$\hat{\dot{X}}_3$
Differentiator
Low Pass Filter

Figure 3 DEGENERATE ESTIMATOR ($K_1 = K_2 = 0$)

THE APPLICATION OF QUADRATIC PROGRAMMING IN ADAPTIVE FILTERING

W. LOCHRIE
Member of the Technical Staff
Autonetics, North American Rockwell
Anaheim, California

and

D. ISAACS
Associate Professor
University of California, Irvine
Irvine, California

ABSTRACT

The problem of designing an adaptive filter for a nonlinear dynamic system is formulated as a constrained estimation problem in which the measurement noise variances are estimated. An adaptive filter mechanization with an imbedded constrained estimation algorithm is then developed. A quadratic programming algorithm is incorporated in the mechanization developed for learning these variance values.

I. INTRODUCTION

System performance degradation caused by errors in the knowledge of system statistics is a practical problem encountered in the implementation of the Kalman filter. In this paper an algorithm is obtained in which measurement data is processed in a manner such that the measurement noise variances are "learned." Shellenbarger [1] formulated the problem of learning statistics of a linear system as an estimation problem. The nonnegativity constraints on the variance estimates were handled by resetting negative estimates to zero. In this paper the problem of learning measurement noise variances in a nonlinear dynamic system is formulated as a constrained estimation problem. Just a rudimentary knowledge of the noise statistics permits both lower and upper bound constraints to be imposed on the variance elements being estimated. This provides a natural application for a mechanization developed in a previous work of the authors [2].

II. PROBLEM FORMULATION

Consider the nonlinear system in which the state vector dynamical equation is

$$\dot{\underline{\zeta}} = \underline{f}(\underline{\zeta}) \tag{1}$$

Assume that observations are made at times $t_1, t_2, \ldots, t_k, \ldots$ and also assume that they are related to the state vector by the nonlinear relation

$$\underline{y}_k = \underline{g}(\underline{\zeta}_k) + \underline{v}_k \tag{2}$$

The subscript k on matrices and vectors denotes evaluation at time t_k. The vector $\underline{v}_k$ represents zero mean white measurement noise with an unknown diagonal covariance matrix R. The assumption of a diagonal covariance implies that the components of the measurement vector are uncorrelated and is valid for many practical systems. The assumption is also made that the state vector estimation error can be described by a linear expansion about the estimated trajectory. This means that the first order Kalman filter [3, 4] can be used in the system mechanization.

III. SYSTEM ANALYSIS

First Order Error Equations

Before deriving the adaptive system algorithm it is necessary to obtain the first order error equations. The tilde is used to distinguish between the actual system state vector, $\underline{\zeta}(t)$, and the filter's estimate of the state vector, $\underline{\tilde{\zeta}}(t)$. The estimated state vector obeys the dynamical equation

$$\dot{\underline{\tilde{\zeta}}}(t) = \underline{f}(\underline{\tilde{\zeta}}) \tag{3}$$

Now define the error vector $\underline{z}(t)$ as follows:

$$\underline{z}(t) \triangleq \underline{\zeta}(t) - \underline{\tilde{\zeta}}(t) \tag{4}$$

Expanding the actual nonlinear system in a Taylor series about $\underline{\tilde{\zeta}}(t)$ and assuming that the second- and higher-order terms in the expansion are negligible yields the first order dynamical equation.

$$\dot{\underline{z}}(t) = F(t)\,\underline{z}(t) \qquad \text{where} \qquad F(t) = \frac{\partial \underline{f}(t)}{\partial \underline{\zeta}(t)}\,. \tag{5}$$

The actual measurement is given by (2) and the expected measurement is given by

$$\underline{\tilde{Y}}_k = \underline{g}(\underline{\tilde{\zeta}}_k) \tag{6}$$

Let $\underline{y}_k$ denote the measurement residual

$$\underline{y}_k \triangleq \underline{Y}_k - \underline{\tilde{Y}}_k \tag{7}$$

This residual can be expressed in terms of the error vector $\underline{z}_k$ by using a Taylor series expansion of (2) truncated after the first-order term

$$\underline{y}_k = M_k\,\underline{z}_k + \underline{v}_k \tag{8}$$

where

$$M_k = \frac{\partial \underline{g}}{\partial \underline{\zeta}} \qquad \text{and} \qquad \operatorname{cov}(\underline{v}_k) = R \tag{9}$$

Linear Filter Model

The first step in designing a linear filter involves modeling the first order error equations given above. It is assumed that the only significant model discrepancy is in the value of the measurement noise covariance matrix. Thus, the filter error model is assumed to have the correct dynamical equation

$$\dot{\underline{x}}(t) = F(t)\,\underline{x}(t) \tag{10}$$

where F(t) is given in (5). The well-known solution to (10) is simply

$$\underline{x}(t) = \theta(t, t_k)\,\underline{x}_k \tag{11}$$

where the transition matrix $\theta(t, t_k)$ is obtained by integrating

$$\dot{\theta}(t, t_k) = F(t)\,\theta(t, t_k) \qquad \text{with} \qquad \theta(t_k, t_k) = I \tag{12}$$

The measurement residual defined in (7) is assumed to be related to the filter's error vector as follows:

$$\underline{y}_k = M_k\,\underline{x}_k + \underline{\tilde{v}}_k \tag{13}$$

where M_k is defined in (9). The vector $\underline{\tilde{v}}_k$ denotes white noise with zero mean and

$$\operatorname{cov}(\underline{\tilde{v}}_k) = \tilde{R}_k \ . \tag{14}$$

First Order Kalman Filter

Assuming that the deviations from the nominal are small enough so that the first-order characterization is valid, then the filter mechanization which is optimal in the sense of minimum mean square error is the first order Kalman filter.

The optimum estimate of $\underline{x}_k$ is the expected value of $\underline{x}_k$ given the measurements $\underline{y}_1, \ldots, \underline{y}_k$; denote this estimate by $\underline{\hat{x}}_{k+}$. Let $\underline{\hat{x}}_{k-}$ denote the expected value of $\underline{x}_k$ given the measurements $\underline{y}_1, \ldots, \underline{y}_{k-1}$. In the mechanization to be described the trajectory, $\underline{\tilde{\zeta}}(t)$, is updated at each measurement as follows:

$$\underline{\tilde{\zeta}}_{k+} = \underline{\tilde{\zeta}}_{k-} + \underline{\hat{x}}_{k+} \tag{15}$$

The vector $\underline{\tilde{\zeta}}_{k-}$ is obtained by integrating the differential equation in (3) from t_{k-1} to t_k with initial condition $\underline{\tilde{\zeta}}_{(k-1)+}$.

The trajectory is reinitialized at each measurement in order to more nearly insure that the linearized model is an adequate description of the system. Note that as a result of this updating, the expected value of the error vector, $\underline{\hat{x}}_{(k+1)-}$, is reset to zero.

The Kalman filter algorithm for computing the optimal estimate $\underline{\hat{x}}_{k+}$ will now be given. First, the estimation error covariance matrix is propogated from t_{k-1} to t_k as follows:

$$P_{k-} = \theta_{(k-1)} P_{(k-1)+} \theta_{(k-1)}^{T} \qquad \text{with} \qquad P_{0+} = E[\underline{x}_0 \ \underline{x}_0^{T}] \tag{16}$$

The measurement residual covariance matrix is then computed.

$$\tilde{C}_k = M_k P_{k-} M_k^{T} + \tilde{R}_k \tag{17}$$

The optimal gain matrix is calculated using the following equation.

$$B_k = P_{k-} M_k^{T} \tilde{C}_k^{-1} \tag{18}$$

The optimal estimate of the error vector is simply

$$\underline{\hat{x}}_{k+} = B_k \underline{y}_k \tag{19}$$

Now for notational convenience, the following definition is made.

$$J_k \triangleq I - B_k M_k \tag{20}$$

The covariance matrix after processing the measurement residual $\underline{y}_k$ is

$$P_{k+} = J_k P_{k-} J_k^{T} + B_k \tilde{R}_k B_k^{T} \tag{21}$$

Error Covariance Matricies

The error covariance expressions which are necessary in the derivation of the adaptive algorithm will be obtained. Denote the error in the estimate of the error vector $\underline{z}_k$ as follows:

$$\underline{q}_{k-} = \underline{z}_k - \hat{\underline{x}}_{k-} \qquad \text{and} \qquad \underline{q}_{k+} = \underline{z}_k - \hat{\underline{x}}_{k+}$$

Next, define the estimation error covariance matricies.

$$Q_{k-} \triangleq E(\underline{q}_{k-}\, \underline{q}_{k-}^{T}) \qquad \text{and} \qquad Q_{k+} \triangleq E(\underline{q}_{k+}\, \underline{q}_{k+}^{T})$$

The propogation of the error covariance is governed by the relation

$$Q_{k-} = \theta_{(k-1)}\, Q_{(k-1)+}\, \theta^{T}_{(k-1)} \tag{22}$$

When the measurement $\underline{y}_k$ is processed the estimation error covariance is modified as follows:

$$Q_{k+} = J_k\, Q_{k-}\, J_k^T + B_k\, R\, B_k^{T} \tag{23}$$

Employing (8) and using the fact that the measurement noise and the estimation error are uncorrelated yields the expression for the measurement residual covariance.

$$C_k = E(\underline{y}_k\, \underline{y}_k^{T}) = M_k\, Q_{k-}\, M_k^{T} + R \tag{24}$$

IV. DERIVATION OF THE ADAPTIVE ALGORITHM

System Equations

It is the purpose of the adaptive algorithm to utilize the measurement data to modify the filter measurement noise variance, $\tilde{R}_k$, in such a manner that it converges to the true measurement noise covariance R. It is assumed that R is a diagonal matrix with unknown variances that are either constant or slowly varying with respect to the time required for $\tilde{R}_k$ to converge to R. Thus, R, which can be viewed as a "state matrix" to be estimated, has the simple dynamical equation

$$\dot{R} = O \text{ (Null Matrix)} \tag{25}$$

It is reasonable to assume that an elementary knowledge of the physical characteristics of the sensors permits both lower and upper bound constraints to be imposed on the noise variance elements being estimated. Letting $\underline{r}$ denote the vector containing the diagonal elements of R and letting $\underline{\alpha}$ and $\underline{\omega}$ be constant vectors this can be expressed as

$$\underline{O} \leqslant \underline{\alpha} \leqslant \underline{r} \leqslant \underline{\omega} \tag{26}$$

The data input to the adaptive system is obtained by using the measurement residual $\underline{y}_k$ to form the sample measurement residual covariance matrix C_k^*. This sample covariance is an approximation to the measurement residual covariance C_k given in (24) and is computed as follows:

$$C_k^{*} = \underline{y}_k\, \underline{y}_k^{T} \tag{27}$$

Denote the difference between the covariance matricies C_k^* and C_k by the noise matrix N_k. That is,

$$N_k = C_k^* - C_k \tag{28}$$

It follows from (24) and (27) that N_k has zero mean. Now the output equation, which relates the state matrix R to the measurement matrix C_k^*, can be obtained by substituting (28) into (24).

$$C_k^* = M_k Q_{k-} M_k^T + R + N_k \tag{29}$$

Model Development

Between measurements the dynamical equation for the estimated state matrix is

$$\dot{\tilde{R}}_k = O \quad \text{(Null Matrix)} \tag{30}$$

The filter's output equation which gives the expected measurement residual covariance matrix is given by (17). As is apparent in equations (16) through (21) $\tilde{R}_k$ propagates into P_{k-} in a nonlinear fashion. Thus, in order to make the problem computationally tractable a linearized model must be developed. To this end, let ΔR_k denote the error in the knowledge of R. That is,

$$\Delta R_k = R - \tilde{R}_k \tag{31}$$

It is apparent from (25) and (30) that between measurements the dynamics for this error matrix is simply

$$\Delta \dot{R}_k = O \tag{32}$$

Now define a residual measurement matrix, ΔC_k, as the difference between the actual measurement matrix C_k^* and the expected measurement matrix $\tilde{C}_k$.

$$\Delta C_k = C_k^* - \tilde{C}_k \tag{33}$$

Finally, represent the difference between the true estimation error covariance and the filter's value of the estimation error covariance as follows:

$$\Delta Q_{k-} = Q_{k-} - P_{k-} \quad \text{and} \quad \Delta Q_{k+} = Q_{k+} - P_{k+} \tag{34}$$

The output model relating the state matrix error to the measurement matrix residual can now be obtained by subtracting (17) from (29).

$$\Delta C_k = M_k \, \Delta Q_{k-} \, M_k^T + \Delta R_k + N_k \tag{35}$$

It is now necessary to get dynamic relations for the matricies ΔQ_{k-} and ΔQ_{k+}. Subtracting (16) from (22) yields the propagation equation

$$\Delta Q_{k-} = \theta_{(k-1)} \Delta Q_{(k-1)+} \theta_{(k-1)}^T \tag{36}$$

The readjustment which occurs after processing the measurement at time t_k is obtained by subtracting (21) from (23).

$$\Delta Q_{k+} = J_k \Delta Q_{k-} J_k + B_k \Delta R_k B_k^T \tag{37}$$

Since the filter's initial value for the estimation error covariance, P_{O+}, is assumed to be correct it is apparent that ΔQ_{O+} is the null matrix.

Note that ΔQ_k is the error in the knowledge of Q_k; the matrix Q_k is the covariance of the error in the estimate of $\underline{z}_k$ and the vector $\underline{z}_k$ is the error in the estimate of $\underline{\zeta}_k$. Thus, although the two linearizations simplify the computations they certainly complicate the linguistics!

Derivation of the Linear Output Relation

It is now necessary to write ΔC_k in a manner such that the elements of ΔR_k can be factored and therefore easily estimated. A clever method of writing matricies which permits this factoring was suggested by Shellenbarger [1]. Denote the diagonal element in the ith row of the matrix ΔR_k by ρ_k. Define a projection matrix U^i such that the diagonal element in the ith row is one and the remaining elements are zeros.

Hence, the m by m measurement noise error covariance matrix can be written as

$$\Delta R_k = \sum_{i=1}^{m} U^i \rho_k^i \tag{38}$$

This representation permits factoring of the various elements, ρ_k^i, when ΔR_k is involved in matrix multiplications. Observation of equations (35), (36) and (37) makes it apparent that it is necessary to get recursive relations for ΔQ_{k-} and ΔQ_{k+} with the ρ_k^i factored. By writing out the first few terms it becomes apparent that

$$\Delta Q_{k-} = \sum_i S_{k-}^i \rho_k^i \qquad \text{and} \qquad \Delta Q_{k+} = \sum_i S_{k+}^i \rho_k^i \tag{39}$$

where

$$S_{k-}^i = \theta_{(k-1)} S_{(k-1)}^i \theta_{(k-1)}^T \quad \text{with} \quad S_{O+}^i = O \tag{40}$$

and

$$S_{k+}^i = J_k S_{k-}^i J_k^T + B_k U^i B_k^T \tag{41}$$

Substituting (38) and (39) into (35) yields an output model in which the measurement residual matrix is a linear function of the variance elements.

$$\Delta C_k = \sum_i (M_k S_{k-}^i M_k^T + U^i) \rho_k^i + N_k \tag{42}$$

Restructuring the Problem

It is now necessary to restructure the problem such that it fits the format given in the development of the first order REAC mechanization [2]. In order to permit a lucid discussion of this development it is necessary to specify m, the number of elements in the measurement vector. Since the example in the next section has a two element measurement vector the value two is used for m.

In order to simplify the notation make the following definition

$$V_k^i \triangleq M_k S_{k-}^i M_k^T \; ; \; i = 1, 2 \tag{43}$$

Hence, equation (42) becomes

$$\Delta C_k = \sum_{i=1}^{2} \left(V_k^{\,i} + U_k^{\,i} \right) \rho_k^{\,i} + N_k \tag{44}$$

This matrix equation will now be converted to a vector equation using a technique suggested by Shellenbarger [1]. Denote the elements of the matricies in (44) as follows:

$$V_k^{\,1} = \{ v_k^{\,1}\ (i,j) \}\ ;\ V_k^{\,2} = \{ v_k^{\,2}\ (i,j) \}$$

and (45)

$$\Delta C_k \triangleq \{ \delta_k(i,j) \}\ ;\ N_k = \{ n_k\ (i,j) \}\ .$$

Noting that ΔC_k is symmetric this output equation can be written in vector form.

$$\underline{\delta}_k = E_k\, \underline{\rho}_k + \underline{n}_k \tag{46}$$

where

$$\underline{\delta}_k = \begin{bmatrix} \delta_k\ (1,1) \\ \delta_k\ (2,2) \\ \delta_k\ (1,2) \end{bmatrix} \ ;\ E_k = \begin{bmatrix} v_k^1\ (1,1) + 1 & v_k^2\ (1,1) \\ v_k^1\ (2,2) & v_k^2\ (2,2) + 1 \\ v_k^1\ (1,2) & v_k^2\ (1,2) \end{bmatrix}$$

and (47)

$$\rho_k = \begin{bmatrix} \rho_k^{\,1} \\ \\ \rho_k^{\,2} \end{bmatrix} \ ;\ \underline{n}_k = \begin{bmatrix} n_k\ (1,1) \\ n_k\ (2,2) \\ n_k\ (1,2) \end{bmatrix}$$

Note that between measurements the dynamical equation for $\underline{\rho}_k$ follows directly from (32) and is simply

$$\dot{\underline{\rho}}_k = \underline{O}\ . \tag{48}$$

Next consider the imposition of the upper and lower bound constraints given in (26). First, however, recall that $\tilde{R}_{k-1}$ is the measurement noise matrix used in the Kalman equations to update the state estimate at time t_{k-1}. Immediately after the state vector $\tilde{\underline{\zeta}}_{(k-1)+}$ is computed, $\tilde{R}_{k-1}$ is updated to $\tilde{R}_k$. The updating equation is simply

$$\tilde{\underline{r}}_k = \tilde{\underline{r}}_{k-1} + \hat{\underline{\rho}}_k \tag{49}$$

where $\tilde{\underline{r}}_j$ denotes the vector containing the diagonal elements of $\tilde{R}_j$ and $\hat{\underline{\rho}}_k$ denotes the estimate of $\underline{\rho}_k$. The estimate $\hat{\underline{\rho}}_k$ is chosen to minimize a mean square error criterion subject to the constraints

$$\underline{O} \leqslant \underline{\alpha} \leqslant \tilde{\underline{r}}_k \leqslant \underline{\omega} \tag{50}$$

Substituting (49) into (50) permits the constraint equation to be written in terms of $\hat{\underline{\rho}}_k$.

$$H \hat{\underline{\rho}}_k \leq \underline{b}_k \tag{51}$$

where

$$H = \begin{bmatrix} I \\ \\ -I \end{bmatrix} \qquad \text{and } \underline{b}_k = \begin{bmatrix} \underline{\omega} - \tilde{\underline{r}}_{k-1} \\ \\ \tilde{\underline{r}}_{k-1} - \underline{\alpha} \end{bmatrix} \tag{52}$$

The problem can now be restated as follows: given that $\underline{\rho}_k$ is governed by the dynamical equation (48) and output equation (46) find the estimate of $\underline{\rho}_k$, denoted by $\hat{\underline{\rho}}_k$, which minimizes the mean square estimation error subject to the constraint equation (51). This formulation permits the first order REAC mechanization to be applied. The algorithm is summarized in the next section.

V. ADAPTIVE SYSTEM ALGORITHM

Standard Portion

1. The sensors provide the measurements at time t_k and both the time and the data vector $\underline{Y}_k$ are stored in the computer.
2. The transition matrix differential equation, given in (12), and the state dynamical equation given in (3) are integrated from t_{k-1} to t_k yielding θ_{k-1} and $\tilde{\underline{\zeta}}_{k-}$ respectively.
3. The expected measurement $\tilde{\underline{Y}}_k$ is computed using (6) and then the measurement residual is formed: $\underline{y}_k = \underline{Y}_k - \tilde{\underline{Y}}_k$.
4. The output matrix M_k is evaluated and then the Kalman equations (16), (17), (18), (19), (20) and (21) are processed. The matricies $\tilde{C}_k$, B_k and J_k are saved for use in the adaptive section of the algorithm. The state vector is updated to $\tilde{\underline{\zeta}}_{k+}$ using (15).

Adaptive Portion

5. Next, the sample covariance is computed: $C_k^* = \underline{y}_k \underline{y}_k^T$. Then the residual covariance matrix is calculated: $\Delta C_k = C_k^* - \tilde{C}_k$. The matricies S_{k-}^i, V_k^i and S_{k+}^i are evaluated in that order using equations (40), (43) and (41) respectively. Using the definitions in (45) and (47) the vector δ_k and matrix E_k are formed from the elements in ΔC_k and V_k.
6. The cost functional in the first order REAC mechanization is updated as follows: $D_k = D_{k-1} + E_k^T Q E_k$, $\underline{c}_k = 2 E_k^T Q \underline{\delta}_k$ and $f(\hat{\underline{\rho}}_k) = -\hat{\underline{\rho}}_k^T D_k \hat{\underline{\rho}}_k + \underline{c}_k^T \hat{\underline{\rho}}_k$
7. It is computationally efficient to first solve the unconstrained problem and then check to see if the constraints are activated. That is, $\hat{\underline{\rho}}_k = 1/2\, D_k^{-1} \underline{c}_k$. (Note that D_k is an m by m matrix where m is just the length of the measurement vector.) Update $\tilde{\underline{r}}_k$ as shown in (49) and check to see if the constraints in (50) are satisfied. If they are satisfied the cycle is complete and if not proceed to step 8.
8. The constraint vector $\underline{b}_k$ is formed as shown in (52) and then the following quadratic programming problem is solved:

   maximize $f(\hat{\underline{\rho}}_k)$ subject to $H \hat{\underline{\rho}}_k \leq \underline{b}_k$.

   The optimal estimate $\hat{\underline{\rho}}_k$ is added to $\tilde{\underline{r}}_{k-1}$ to form $\tilde{\underline{r}}_k$.

VI. ADAPTIVE SATELLITE NAVIGATION SYSTEM EXAMPLE

A precision autonomous satellite navigation system provides an excellent practical example for demonstrating the adaptive algorithm discussed in the previous section. The system considered provides a continuous estimate of the satellite's orbit and periodically corrects this estimate using data provided by two sensors onboard the satellite. One of the sensors, a radar altimeter, measures the satellite altitude. The other sensor a horizon scanner detects the thermal discontinuity between the earth's horizon and the space background. This observation in essence provides a measurement of the local vertical.

In the Monte Carlo simulation of the navigation system the satellite was assumed to be in a nearly circular 100 nm orbit. Radar altimeter and horizon scanner measurements were simulated every 45 seconds for the 60 minute run. The "actual" measurement noise values were obtained using the IBM gaussian random number generator. The noise values generated for each sensor had essentially zero mean and the standard deviations were 1825 feet for the radar altimeter and 228 $\widehat{\sec}$ for the horizon scanner.

It was assumed that testing of the sensors combined with an analysis of their mechanization permitted the following constraints to be placed on the noise variance values.

$$\begin{bmatrix} (500\ \text{ft})^2 \\ (60\ \widehat{\sec})^2 \end{bmatrix} \leqslant \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \leqslant \begin{bmatrix} (2700\ \text{ft})^2 \\ (330\ \widehat{\sec})^2 \end{bmatrix}$$

Results which demonstrate the variance learning capability of the system are shown in Figure 1. The two runs shown demonstrate that convergence occurs for a broad range of initial conditions. It should be noted that one Monte Carlo run, which consisted of cycling through the adaptive algorithm 80 times took only 6 seconds on the IBM 360/91 Computer. A modified version of Wolfe's quadratic programming method [5] was used in the algorithm. More details concerning this example can be found in [6].

VII. SUMMARY

In this paper an algorithm for implementing an adaptive first order Kalman filter has been presented. The term adaptive is used in the sense that the filter's value for the measurement noise covariance is modified by making use of the measurement data; that is, the filter is operated in "closed loop" fashion. This problem was formulated as a first order constrained estimation problem imbedded in a first order Kalman filter. A Monte Carlo simulation of an adaptive satellite navigation system was used to successfully demonstrate this algorithm.

REFERENCES

[1] Shellenbarger, J. C., "A Multivariance Learning Technique for Improved Dynamic System Performance," National Electronics Conference, October, 1967.

[2] Isaacs, D. and W. Lochrie, "Estimation and Identification of Nonlinear Systems with Constraints," Seventh Annual Allerton Conference on Circuit and Systems Theory, October, 1969.

[3] Smith, G. L., S. F. Schmidt, and L. A. McGee, "Application of Statistical Filter Theory to the Optimal Estimation of Position and Velocity on board a Circumlunar Vehicle," NASA TR-R135, '62.

[4] Gunckel, T. L., "Orbit Determination Using Kalman's Method," Vol. 10, No. 3, pp. 273-291, Journal of the Institute of Navigation, Autumn, 1963.

[5] Wolfe, P., "The Simplex Method for Quadratic Programming," Econometrica, 27, 1959, pp. 382-398.

[6] Lochrie, W.D., "The Application of Nonlinear Programming to Dynamic System Estimation, Identification and Adaptation," PhD Dissertation, University of California, Irvine; June, 1970.

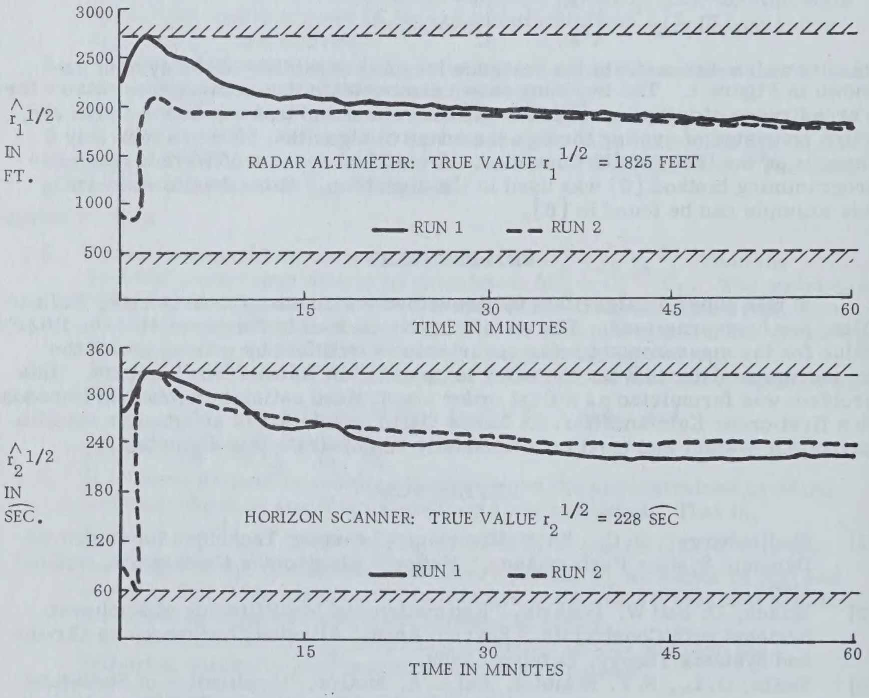


Figure 1. Results of the Adaptive System Simulation

CONSTRAINED OPTIMIZATION OF NONLINEAR FILTERS AND SYSTEMS ON AN APPROPRIATE SYSTEM SPACE

N. U. AHMED[*]
University of Ottawa,
Ottawa, Canada.

ABSTRACT

In this paper the concept of a linear normed "System Space" $G_\lambda(\lambda > 0)$ is introduced. It is shown that $G_\lambda(\lambda > 0)$ is a Hilbert space and that it satisfies many of the known properties of an abstract Hilbert space.

The class of nonlinear functionals based on the Wiener process $\{x(t), t \in I\}$ with $\mathcal{E}(x(t) \cdot x(s)) = \frac{1}{2\lambda} \min(t, s)$ and also stochastic hereditary (or non hereditary) differential equations of Ito type are represented by a characteristic element of the "Systems-Space" G_λ.

The criterion for weak and strong compactness of subsets of the system space G_λ are presented. These criterion are found very useful in the synthesis of nonlinear systems with constraints on the system space. This is illustrated by an example. Extension of these results to more general system spaces G_λ^p $(p \geq 1)$ $\lambda > 0$ is briefly indicated.

INTRODUCTION

Based on the earlier results of Wiener (1), recently Brick (2) and the author (3,4) reported some interesting results in the theory of nonlinear systems. These results are quite useful in the Fourier analysis of nonlinear systems defined on the Wiener measure space (Ω, B, μ_λ). Recently it was reported in (5) that the solution of a large class of stochastic (hereditary or non hereditary) Ito differential equations (5,6) has a canonical representation with random Fourier Kernels. These results are useful in the analysis and synthesis of nonlinear systems provided there is no constraint.

In this paper we are interested in the problem of synthesis of nonlinear systems and filters with constraints. For this purpose we introduce the normed vector space G_λ(for $\lambda > 0$) defined by

$$G_\lambda = \{K \triangleq (K_0, K_1, \ldots K_s \ldots) : K_s \in L^2(I^s) \text{ for each } s \in N$$

$$\text{and that } \quad \| K \|^2_\lambda \triangleq \sum_{s=o}^{\infty} \frac{s!}{(2\lambda)s} \int \cdots \int_{I^s} | K_s(\tau_1, \ldots, \tau_s) |^2 d\tau_1 \ldots d\tau_s < \infty \}$$

where I^s is the s-copies of the interval I on the real line, λ is any positive real number, N is the set of non-negative integers and, for $s = 0$, $L^2(I^o)$ stands for the field of scalars.

In section 2 certain basic properties of the space G_λ are developed and it is shown there that for each $\lambda > 0$ G_λ is a Hilbert space. In section 3 necessary and sufficient conditions for weak and strong compactness of a subset F of the "System-Space" G_λ are presented.

---

[*] This work was supported in part by the National Research Council of Canada under grant No A-7109.

These results combined with a recent result (5, prop 3) are then used in section 4 to obtain a synthesis (in the weak sense) of a large class of nonlinear stochastic differential systems of Ito type with constraints on G_λ.

Possible extensions of these results to the more general class of "Systems-Spaces" G^p_λ $(p \geq 1)$ are briefly indicated in section 5.

2. SOME BASIC PROPERTIES OF THE SPACE G_λ.

In the following we will observe that G_λ is a Hilbert space.

Definition 1.

Consider the ampping that takes $G_\lambda X\, G_\lambda$ into C - the field of scalars - by the rule $\{K, L\}$ $\sum_{s=o}^{\infty} \frac{s!}{(2\lambda)^s} (K_s, L_s) \triangleq (K,L)_\lambda$ where (K_s, L_s) is the usual innerproduct defined for $L^2(I^s)$ spaces and $\|K_s\|$ is the norm of $K_s \in L^2(I^s)$.

Lemma 1.

For all K and $L \in G_\lambda$ $|(K,L)_\lambda| \leq \|K\|_\lambda \; \|L\|_\lambda$ 2.1

$$\text{where} \quad \|K\|_\lambda \triangleq \left(\sum_{s=o}^{\infty} \frac{s!}{(2\lambda)^s} \|K_s\|^2 \right)^{1/2} .$$

Lemma 2.

For all K and $L \in G_\lambda$ and $a \in C$; $K + L \in G_\lambda$, $a\,K \in G_\lambda$ and that

$\|K + L\|_\lambda \leq \|K\|_\lambda + \|L\|_\lambda$. 2.2

It follows from the above Lemmas that G_λ is a linear vector space over C with an innerproduct satisfying the following properties:

For all K, L, $M \in G_\lambda$ and $a \in C$

$$(K + L, M)_\lambda = (K,M)_\lambda + (L,M)_\lambda$$
$$(a\,K, L)_\lambda = a\,(K,L)_\lambda$$
$$(K, L)_\lambda = (L, K)^*_\lambda$$
$$\|K\|^2_\lambda = (K,K)_\lambda \quad > \quad 0 \;\; \text{if}\;\; K \neq 0 \in G_\lambda.$$

where z^* is the complex conjugate of the complex number z. The above relations imply $(0,K)_\lambda = (K,0)_\lambda = 0$; $(K, aL)_\lambda = a^*(K,L)_\lambda$ and $(K, L+M)_\lambda = (K,L)_\lambda + (K,M)_\lambda$.

Remark: A sequence in G_λ will be denoted by using a superscript $\{K^n\} \triangleq \{(K^n_o, K^n_1 \ldots K^n_s \ldots)\} \in G_\lambda$.

Proposition 1.

G_λ is a Hilbert space.

Proof: It follows from the above discussion that G_λ is an innerproduct space so all that is required to prove is that G_λ is a Banach space (a complete normed linear space) with respect to its norm topology as defined above. Let $\{K^n\} \in G_\lambda$ be a cauchy sequence then by the corresponding definition we have

$$\lim_{m,n \quad \infty} \| K^m - K^n \|_\lambda^2 = \lim_{m,n \quad \infty} \sum_{s=o}^{\infty} \frac{s!}{(2\lambda)^s} \| K_s^m - K_s^n \|^2 = 0.$$

This implies that for each $s \in N$, $\{K_s^n\}_n \in L^2(I^s)$ is a cauchy sequence and since $L^2(I^s)$ is a Banach space $K_s^n \quad K_s^o$ in the norm topology of $L^2(I^s)$ and that $K_s^o \in L^2(I^s)$. Let us define $K^o \underline{\triangle} (K_o^o, K_1^o, K_2^o \ldots K_s^o \ldots)$. We must show that $K^n \quad K^o$ in the norm topology of G_λ and that $K^o \in G_\lambda$. Since for each $s \in N$, $K_s^r \quad K_s^o$, $\underline{\lim}_r \| K_s^r - K_s^o \| = 0$ and therefore it follows from the inequality $\| K_s^o - K_s^n \| \leq \| K_s^o - K_s^r \| + \| K_s^r - K_s^n \|$ that $\| K_s^o - K_s^n \| \leq \underline{\lim}_r \| K_s^r - K_s^n \|$. Define $g^r = K_s^o - K_s^r$ and $h^r = K_s^r - K_s^n$. It is clear that $g^r \rightarrow o \in L^2(I^s)$ and $h^r \rightarrow K_s^o - K_s^n \in L^2(I^s)$. From the continuity of innerproduct in L^2 spaces it follows that

$$\lim_{r \rightarrow \infty} (g^r, h^r) = (0, K_s^o - K_s^n) = 0.$$

Hence it follows from the equality

$$\| K_s^o - K_s^n \|^2 = \| K_s^o - K_s^r \|^2 + \| K_s^r - K_s^n \| + 2 R_e (K_s^o - K_s^r, K_s^r - K_s^n)$$

that $\| K_s^o - K_s^n \|^2 \leq \underline{\lim}_r \| K_s^r - K_s^n \|^2$ also. Thus

$$\sum_{s=o}^{\infty} \frac{s!}{(2\lambda)^s} \| K_s^o - K_s^n \|^2 \leq \sum_{s=o}^{\infty} \frac{s!}{(2\lambda)^s} \underline{\lim}_r \| K_s^r - K_s^n \|^2. \qquad 2.3$$

It may be interesting to notice that Fatou's lemma (7, Dunford and Schartz, pp 153) originally proved for non-negative measurable functions also applies to infinite series of non-negative real numbers.

Therefore we have,

$$\sum_{s=o}^{\infty} \frac{s!}{(2\lambda)^s} \underline{\lim}_r \| K_s^r - K_s^n \|^2 \leq \underline{\lim}_r \sum_{s=o}^{\infty} \frac{s!}{(2\lambda)^s} \| K_s^r - K_s^n \|^2. \qquad 2.4$$

From the inequalities 2.3 and 2.4 and the definition of the elements K^o and K^n we have $\| K^o - K^n \|_\lambda \leq \underline{\lim}_r \| K^r - K^n \|$. 2.4'

Since $\{K^n\} \in G_\lambda$ is a cauchy sequence it follows from 2.4' that for every $\epsilon > o$ there exists a natural number $n_o(\epsilon)$ such that $\underline{\lim}_r \| K^r - K^n \|_\lambda \leq \epsilon$ for all $n \geq n_o(\epsilon)$. Thus $\| K^o - K^n \|_\lambda \leq \epsilon$ for all $n \geq n_o(\epsilon)$ ie $K^n \rightarrow K^o$ in the norm topology of G_λ. Further $K^o = K^o - K^n + K^n$ and therefore $\| K^o \|_\lambda \leq \epsilon + \| K^n \|_\lambda$ for $n \geq n_o(\epsilon)$ and since $K^{n_o} \in G_\lambda$, we have $K^o \in G_\lambda$. Thus G_λ is a Banach space. This completes the proof of the proposition.

Lemma 3.

The following equality known as the parallelogram identity

$$\|K + L\|_\lambda^2 + \|K - L\|_\lambda^2 = 2\|K\|_\lambda^2 + 2\|L\|_\lambda^2 \qquad 2.5$$

holds for all $K, L \in G_\lambda$.

We will need the following propositions in the sequel.

Proposition 2.

Let S be a nonempty closed convex subset of G_λ and suppose $d = \inf_{K \in S} \|K\|_\lambda$. Then there exists a unique $K^o \in S$ such that $\|K^o\|_\lambda = d$.

Proof: see (Ref 13)

The following definition will be useful in the sequel.

Definition 2.

The space conjugate to G_λ and denoted by G_λ^* is the space of all bounded linear functionals on G_λ.

Proposition 3.

Let H be a hyperplane in G_λ defined as $H = \{K \in G_\lambda : L^*(K) = 1\}$ where $L^* \in G_\lambda^*$. Then there exists a unique $K^o \in H$ such that $(M, K^o)_\lambda = 0$ whenever $L^*(M) = 0$ for certain $M \in G_\lambda$.

Proof: see (Ref 13)

Proposition 4.

To each $L^* \in G_\lambda^*$ - bounded linear functionals on G_λ - there corresponds a unique $L \in G_\lambda$ such that $L^*(K) = (K, L)_\lambda$ for all $K \in G_\lambda$.

Proof: (see Ref 13)

3. STRONGLY AND WEAKLY COMPACT SUBSETS OF G_λ

We are now prepared to study the question of compactness of subsets of G_λ. The following definitions are standard.

Definition 3.

A sequence $\{K^n\} \in G_\lambda$ is said to be weakly convergent to an element $K^o \in G_\lambda$ if $\lim_n L^*(K^n) = L^*(K^o)$ for every $L^* \in G_\lambda^*$.

Definition 4.

(i) A set $F \subset G_\lambda$ is said to be sequentially compact if every infinite sequence from F has a convergent subsequence converging in the norm topology to an element in G_λ and it is said to be closed if the limit also belongs to F.

(ii) A set $F \subset G_\lambda$ is said to be weakly sequentially compact if every infinite sequence from F has a weakly convergent subsequence. Further F is said to be weakly closed if this limit also belongs to F.

Definition 5.

A set $F \subset G_\lambda$ is said to be bounded if $\sup_{K \in F} \| K \|_\lambda \leq a$ for some some finite real number $a \geq 0$.

Definition 6.

A Banach space B is said to be uniformly convex if for every $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that for $x, y \in B$, $\| x \| \leq 1$, $\| y \| \leq 1$, and $\| x-y \| \geq \epsilon$ imply that $\| x+y \| \leq 2(1 - \delta(\epsilon))$.

Proposition 5

A set $F \subset G_\lambda$ is weakly (sequentially) compact if, and only if, it is bounded.

Proof: Since the space G_λ satisfies the parallelogram identity (lemma 3) it follows from the definition 6 that the space G_λ is uniformly convex.

Further, the well known Millman's theorem (Yosida, pp 126) states that every uniformly convex Banach space is reflexive (regular). In the case of reflexive Banach spaces a bounded set is weakly compact (Dunford 7, pp 68, th 28).

Therefore F, being a bounded subset of the uniformly convex Banach space G_λ, is weakly compact.

The only if part is almost immediate.

In this section a set of necessary and sufficient conditions for strong compactness of subsets of G_λ are presented.

Let us define a linear operator A^t, $t \geq 0$ on G_λ by $A^t K = (K_o, K_1(t+\tau): \ldots, K_s(t+\tau_1, \ldots t+\tau_s), \ldots)$. It is clear that $K \in G_\lambda$ implies that $A^t K \in G_\lambda$ for all $t \geq o$ and that $\| A^t \| \leq 1$. To avoid additional complications we will assume in this section that the interval I is bounded. Without any loss of generality we may assume, for the sake of simplicity, that for each integer $n = K_n(t+\tau_1, \ldots, t+\tau_n) \equiv 0$ whenever $(t+\tau_1, t+\tau_n) \notin I^n$.

Proposition 6.

A set $F \subset G_\lambda$ is (strongly) conditionally compact if, and only if,

(i) F is bounded:

(ii) $\lim_{n \to \infty} \sum_{s=n}^{\infty} \frac{s!}{(2\lambda)^s} \| K_s \|^2 = 0$ uniformly with respect to $K \in F$,

and (iii) $\lim_{t \to o} \| A^t K - K \|_\lambda = 0$ uniformly with respect to $K \in F$.

Proof: (see Ref 13)

Remarks:

(i) Let, for $\lambda > 0$, (Ω, B, μ_λ) be the Wiener measure space corresponding to the Wiener process $\{x(t), t \in I\}$ with $\mathcal{E}(x(t)x(s)) = \frac{1}{2\lambda} \min(t, s)$. It is known (3,4) that the space G_λ is isometrically isomorphic to

$L^2(\Omega, B, \mu_\lambda)$. This relationship between G_λ and $L^2(\Omega, B, \mu_\lambda)$ is very useful in presenting similar criterion for strong and weak compactness of subsets $L^2(\Omega, B, \mu_\lambda)$.

(ii) When (Ω, B, μ_λ) is a general measure space, for example the Wiener measure space, it is not immediately clear how the criterion given in the theorems 20 and 21 (7, pp 298-301, th 20, 21) are changed. The difficulty lies in the fact that Ω is not a finite dimensional Euclidean space; it is in fact the space of Brownian motion functions. So far as weak compactness is concerned the criterion is identical both for G_λ and any other abstract Hilbert space.

The difference lies in the criterion for strong compactness. The condition (ii) of proposition 6 was found to be a necessary condition for a subset of G_λ to be compact but this is not required for L_2 spaces over any finite dimensional space.

If the basic set I on which the Wiener process is defined is the whole real line then like theorems 20, 21(7, pp 298-301) we must add a fourth condition to proposition 6 giving

$$\lim_{m\to\infty} \sum_{s=0}^{\infty} \frac{s!}{(2\lambda)^s} \int\int_{R^s \setminus I^s_m} | K_s(\tau_1, \dots \tau_s) |^2 \, d\tau_1 \dots d\tau_s = 0 \qquad 3.4$$

uniformly with respect to $K = (K_0, K_1 \dots K_s \dots) \in F$, where $I^s_m \subset R^s$ and $\lim_{m\to\infty} I^s_m \uparrow R^s$; R^s being the s-copies of the real line.

The criterion for compactness of subsets of $L^2(\Omega, B, \mu_\lambda)$ can be stated directly in terms of (a) the ergodic group (1, pp 44) $\{ T^t, t \geq 0 \}$ of measure preserving transformations of the σ-field B into itself and (b) projections in the $L^2(\Omega, B, \mu_\lambda)$ space. Let $\{ H_n(K_n, \sigma), K_n \in L^2(I^n)$; $n = 0, 1, 2 \dots \}$ be the modified Wiener's orthogonal set (1; 4, pp 141) defined by

$$H_n(K_n, \sigma) = \int_{I^n} \dots \int K_n(\tau_1, \dots, \tau_n) \sum_{m=0}^{[\frac{n}{2}]} \frac{n!}{(4\lambda)^m (n-m)! \, m!} S_m(dx(\tau_1, \sigma) \dots dx(\tau_n, \sigma)) \qquad 3.5$$

Where

$$S_m(dx(\tau_1, \sigma), \dots dx(\tau_n, \sigma)) = \prod_{i=1}^{n} dx(\tau_i, \sigma) \qquad \text{for } m = 0$$

$$= \prod_{i=1}^{m} \delta(\tau_i - \tau_{2m+1-i}) \, d\tau_i \, d\tau_{2m+1-i} \prod_{j=2m+1}^{n} dx(\tau_j, \sigma) \text{ for } m \geq 1$$

and $\delta(t)$ is the Dirac delta function.

Let for each $n \in N$, $M^n \triangleq \overline{sp} \{ H_s(K_s, \sigma) : K_s \in L^2(I^s) \; s \geq n \}$

be the closure of the linear subspace of the $L^2(\Omega, B, \mu_\lambda)$ space spanned by the elements $\{H_s : K_s \in L^2(I^s)\ s \geq n\}$ and all their linear combinations and let Q^n be the corresponding projection so that $Q^n f \in M^n$ for each $f \in L^2(\Omega, B, u_\lambda)$. With this preparation we are now in a position to present the criterion for strong conditional compactness of a subset $F \subset L^2(\Omega, B, \mu_\lambda)$ directly.

Corollary 1.

A set $F \subset L^2(\Omega, B, \mu_\lambda)$ is strongly conditionally compact if, and only if

(i) F is bounded

(ii) $\lim\limits_{n \to \infty} \int_\Omega |Q^n f(\sigma)|^2 \, d\mu_\lambda(\sigma) = 0$ uniformly with respect to $f \in F$

and (iii) $\lim\limits_{t \downarrow 0} \int_\Omega | f(T^t \sigma) - f(\sigma) - f(\sigma) |^2 \, d\mu_\lambda(\sigma) = 0$ uniformly with respect to $f \in F$; where $\{Q^n, n \geq o\}$ and $\{T^t, t \geq o\}$ are as defined before.

Proof: The proof follows from proposition 6 and the fact (3,4) that the space $L^2(\Omega, B, \mu_\lambda)$ is isometrically isomorphic to the space G_λ and that compactness and weak compactness are preserved under isomorphism.

Corollary 2.

A set $F \subset L^2(\Omega, B, \mu_\lambda)$ is weakly (sequentially) compact if and only if F is bounded.

Proof: The proof follows from proposition 5 and the fact that $L^2(\Omega, B, \mu_\lambda)$ is isomorphic to G_λ.

Definition 7.

A function φ mapping a Banach space B into the reals is said to be weakly lower-semicontinuous (upper-semicontinuous) at the point $f_o \in B$ if $\varphi(f_o) \leq \underset{n}{\underline{\lim}}\ \varphi(f_n)$ ($\varphi(f_o) \geq \overline{\mathrm{lm}}\ \varphi(f_n)$) for every sequence $\{f_n\} \in B$ that converges weakly to $f_o \in B$. The function φ is said to be weakly lower-semicontinuous (upper-semicontinuous) on a set $F_o \subset B$ if it is so at every point of F_o.

The following results are quoted from M.M. Vainberg (9).

Proposition 7.

A weakly lower-semicontinuous real valued function φ defined on a weakly compact subset F of a Banach space B attains its infimum in F provided it is also weakly closed (Vainberg, pp 78, th 9.2).

Proposition 8.

A weakly upper-semicontinuous real valued function φ defined on a weakly compact subset F of a Banach space B attains it supremum in F provided F is also weakly closed (Vainberg, pp 100, th 13.1).

Proposition 9.

A continuous real valued function φ defined on a subset F of a Banach space B attains both the minimum and maximum in F if F is weakly compact and weakly closed (Vainberg, pp 100 th 13.2).

4. AN APPLICATION

The following example is intended to illustrate a possible application of the space G_λ and its properties to the problems of optimal synthesis of nonlinear filters with constraints.

Consider the following class of stochastic Ito differential equation

$$dy = f\,dt + g\,dx, \quad t \in I \triangleq [0,T] \qquad 4.1$$

where f may take any one of the two forms

(i) $f = f(t,y(t))$ and (ii) $f = f(t, y(s);\ 0 \le s \le t)$ and

Similarly g may take any of the following forms:

(i) $g = g(t,y(t))$, (ii) $g = g(t, y(t), x(t))$, (iii) $g = g(t,y(s);\ 0 \le s \le t)$

or (iv) $g = g(t,y(s), x(s);\ 0 \le s \le t)$.

$y(o) = y_o$ is a given random variable with finite second moment and (Ω_o, B_o, P_o) is the corresponding probability space. $x \triangleq \{x(t),\ t \in I\}$ is the wiener process with $\mathcal{E}(x(t), x(s)) = \frac{1}{2\lambda} \min(t,s)$ $t, s \in I$ and (Ω, B, μ_λ) is the corresponding probability (measure) space where $\lambda > 0$. It is known that if f and g satisfy certain appropriate Lipschitz conditions (5, 6, 10, 11) and if the given initial random variable y_o has bounded second moment then the system 4.1 has a unique (in the stochastic sense) solution y with bounded second moment for each $t \in I$. For convenience of description the solution y may be denoted by $y(t) = y(t, y_o, \sigma), t \in I, y_o \in \Omega_o, \sigma \in \Omega_t$ where Ω_t contained in Ω represents the section $\{x(s)\ 0 \le s \le t\}$ of the wiener process $\{x(s),\ 0 \le s \le T\}$ and B_t contained in B represent the corresponding σ-algebra. The problem is to synthesize a nonlinear system of the class $z(t, \sigma) = H(t, x(s, \sigma);\ 0 \le s \le t)$ so that the output z of this system at time T approximates that of the system 4.1 subject to the constraint that $\int_\Omega |z(T,\sigma)|^2 d\mu_\lambda$ is bounded. Precisely we are interested to find a system H so that its output z at time T minimizes the integral $\int_{\Omega_o} \int_\Omega |z(T,\sigma) - y(T, y_o, \sigma)|^2 d\mu_\lambda\, dP_o$ subject to the constraint that $\int_\Omega |z(T,\sigma)|^2 d\mu_\lambda \le \alpha < \infty$. Without any loss of generality we may assume that $\int_\Omega |z(T,\sigma)|^2 d\mu_\lambda \le 1$. It is known that the process y has a canonical representation (5 prop. 3) and consequently there exists a $L^{y_o} \in G_\lambda$ P_o - a.e so that

$$y(T, y_o, \sigma) \leftrightarrow L^{y_o} \quad \text{where}$$

$$L^{y_o} \triangleq (L_o(y_o/T), L_1(y_o/T, \tau_1), \ldots L_s(y_o/T, \tau_1, \ldots, \tau_s) \ldots).$$

By assumption the filter H is to be chosen so that its output $z(T,\sigma)$ is B-measurable and belongs to $L^2(\Omega, B, \mu_\lambda)$. Therefore corresponding to $z(T,\sigma)$ there is a $K \in G_\lambda$; consequently the problem of optimal synthesis of the filter H can be reformulated in the context of the space G_λ. Find a $K \in G_\lambda$ so that $\int_{\Omega_o} \| K - L^{y_o} \|_\lambda^2 \, dP_o$ is a minimum subject to the constraint that $\|K\|_\lambda \le 1$. Let $S_1 = \{K \in G_\lambda : \|K\|_\lambda \le 1\}$.

Define $\varphi(K) \triangleq \int_\Omega \| K-L^{y_o} \|_y^2 \; dP_o$ for $K \in S_1$. Since S_1 is bounded, by proposition 5, it is weakly compact. S_1 is also weakly closed. The existence of an optimal system H^* subject to the constraint discussed above will follow if we can prove the existence of an element $K^* \in S_1 \subset G_\lambda$ that minimizes the function $\varphi(K)$. By proposition 5 $S_1 \subset G_\lambda$ is weakly compact and therefore, by proposition 7, it follows that if we can show that $\varphi(K)$ is weakly lower semicontinuous on S_1 then there is a $K^* \in S_1$ that minimizes $\varphi(K)$ on S_1.

Let $\{K^n\}$ be any sequence in S_1 converging weakly to $K^o \in S_1$. Then P_o - a.e

$$\| K^n - L^{y_o} \|_\lambda^2 = \| K^n - K^o + K^o - L^{y_o} \|_\lambda^2$$

$$= \| K^n - K^o \|_\lambda^2 + 2(K^n - K^o, K^o - L^{y_o})_\lambda + \| K^o - L^{y_o} \|_\lambda^2$$

$$\geq \| K^o - L^{y_o} \|_\lambda^2 + 2(K^n - K^o, K^o - L^{y_o})_\lambda,$$

and

$$\underline{\lim}_{n} \| K^n - L^{y_o} \|_\lambda^2 \geq \| K^o - L^{y_o} \|_\lambda^2 + 2 \, \underline{\lim}_{n} (K^n - K^o, K^o - L^{y_o})_\lambda \qquad 4.2$$

Since $K^n \to K^o$ weakly, $(K^n - K^o, K^o - L^{y_o})_\lambda \to o$ P_o - a.e.

Thus

$$\underline{\lim}_{n} \| K^n - L^{y_o} \|_\lambda^2 \geq \| K^o - L^{y_o} \|_\lambda^2 \quad P_o \text{ - a.e.} \qquad 4.3$$

Integrating either side of 4.3 we have

$$\varphi(K^o) \leq \underline{\lim}_{n} \varphi(K^n) \qquad 4.4$$

Thus it follows from the definition 7 that the functional φ is weakly lower semicontinuous. Therefore there exists a $K^* \in S_1$ and hence an optimal system H^* (subject to the constraint discussed before) so that $\varphi(K^*) \leq \varphi(K)$ for all $K \in S_1$ and H^* is a weak synthesis of the system 4.1.

Remark:

If $\varphi : G_\lambda \to R$ is continuous on $F \subset G_\lambda$ and F is a strongly compact subset of G_λ then there is both a strong maximum and minimum of φ on F. The corresponding element $K^* \in F$ realizes a strong synthesis H^* for the system 1.4. Since strong compactness always implies weak compactness the strong extremal of φ on F coincides with its weak extremal.

5. EXTENSIONS

Define the normed vector space G_λ^p $(p \geq 1)$ for $\lambda > 0$ by

$G_\lambda^p = \{ K = (K_o, K_1 \ldots K_s \ldots) : K_s \in L^p(I^s)$ for each $s \in N$ and

that $\| K \|_\lambda \triangleq \left(\sum_{s=o}^{\infty} \frac{s!}{(2\lambda)^s} \int \cdots \int_{I^s} | K_s(\tau_1, \ldots \tau_s) |^p \, d\tau_1 \ldots d\tau_s \right)^{1/p} < \infty \}$.

For $p = 2$ we have the Hilbert space G_λ with which this paper was concerned. For $p \geq 1$ it can be shown that G_λ^p is a Banach space and like

L^p spaces over finite dimensional spaces it is possible to show that G^q_λ (where $\frac{1}{p} + \frac{1}{q} = 1$) is the Banach space dual to the Banach space G^p_λ. For $p \geq 2$ it can be shown that the Banach spaces G^p_λ are reflexive and therefore every bounded set in it is weakly compact.

CONCLUSION

In this paper the concept of a normed linear system space G_λ was introduced and certain basic properties of the space G_λ was presented. It was shown that G_λ is a Hilbert space. Certain criterion of weak and strong compactness of a set $F \subset G_\lambda$ were presented. In other words criterion of weak and strong compactness of a family of nonlinear systems in the system space G_λ were presented. This leads to the solution of the problem of weak and strong synthesis of the nonlinear systems with constraints on the system space.

The above result was illustrated by an example from filter theory and stochastic Ito differential equations. It is interesting to note that the system space G_λ includes both hereditary and ordinary stochastic differential systems. Extension to more general system spaces G^p_λ $(p \geq 1)$, $\lambda > 0$ was briefly indicated.

REFERENCES

1. N. Wiener, "Nonlinear problems in Random Theory", Cambridge, Mass. M.I.T. Press, 1958.
2. D.B. Brick, "On the applicability of Wieners canonical expansions" I.E.E.E. Trans. Systems Science and Cybernetics, Vol SSC-4, (March 1968), pp 29-38.
3. N.U. Ahmed, "Fourier analysis on Wiener measure space", J. Franklin Inst. Vol. 286, No. 2,(August, 1968), pp 143-151).
4. N.U. Ahmed "Some novel properties of Wiener's canonical expansion", I.E.E.E. Trans. Systems Science and Cybernetics, Vol. SSC-5 (April 1969), pp 140-144).
5. N.U. Ahmed "Closure and completeness of Wiener's orthogonal set $\{G_n\}$ in the class $L^2(\Omega, B, \mu)$ and its application to stochastic Heriditary Differential Systems" (To appear in Information and Control, August 1970)
6. K. Ito, "On stochastic differential equations" Mem. Am. Math. Soc. No. 4, (1951), pp 29-36.
7. N. Dunford and J.T. Schwartz "Linear Operators" part I, Interscience, 1964.
8. L.V. Kantorovich and G.P. Akilov "Functional analysis in Normed spaces" Eng. Trans. MacMillan, 1964.
9. M.M. Vainberg, "Variational methods for the study of Nonlinear operators" Eng. Trans. Holden Day, 1964.
10. W.H. Fleming and M. Nisio, "On the existence of optimal stochastic controls" Jr. of Math. Mech. Vol. 15, No. 5(1966) pp 777-794.
11. N.U. Ahmed, "Optimal control of a large class of stochastic differential systems with finite memory", 7th. An. Allerton Conf. on Circuit and Systems Theory (1969) pp 44-53.
12. K. Yosida, "Functional Analysis"(Second ed. 1968) pp 126.
13. N.U. Ahmed, "Criterion of Strong and Weak compactness of a family of nonlinear functionals based on the Wiener process and their application to constrained optimization of nonlinear filters" TR 69-8, Oct 1969, Ottawa University.

ON OPTIMAL NONLINEAR ESTIMATION - PART I: CONTINUOUS OBSERVATION

JAMES T. LO

Department of Operations Research, Stanford University, Stanford, Calif.

ABSTRACT

A generalization of the Bucy representation theorem, which plays the same role in the case of general optimal estimation for an arbitrary Markov process as does the Bucy theorem in the case of optimal filtering for a diffusion process, is obtained under very weak hypotheses. It forms the cornerstone of the paper, from which all the other results are deduced.

Some properties of optimal smoothing and prediction are briefly discussed. The conditional densities of the smoothing and prediction processes are expressed in terms of the conditional density of the filtering process.

Smoothing and prediction for the class of systems which reduce, via the concept of finite dimensional sensor orbits, to the case of linear dynamics are investigated, thereby illustrating the applicability of previous results in order to obtain efficiently computable solutions.

1. INTRODUCTION

This paper is primarily concerned with the optimal estimation for an arbitrary continuous or discrete time Markov process observed in white Gaussian noise. The observation process is described, in the sense of K. Ito, by the following stochastic differential equation

$$dz = h(x_t, t)dt + d\nu_t \,, \tag{1.1}$$

where ν_t is a zero mean Brownian motion with covariance

$$\mathcal{E}(\nu_t - \nu_{t_o})(\nu_s - \nu_{t_o})' = \int_{t_o}^{\min(t,s)} R_\tau d\tau \,, \tag{1.2}$$

R_τ being a continuous function of τ. For a discrete time Markov process $x. = \{x_{t_i}\}_i$, the observation is taken to its continuous time version which is defined by $x_t = x_{t_i}$ for $t_i \leq t < t_i$.

The fundamental result of this paper is a generalization of the Bucy representation theorem, which was published in BUCY [1965]. In a recent thesis, LO [1969], it was pointed out that the Bucy theorem played a fundamental role in optimal nonlinear filtering theory, not only in the sense that the stochastic differential equations for the conditional probability density and the conditional moments may be easily derived from it, but also that we may derive some effectively computable formulas for the optimal filtering directly from the theorem. [We note that the stochastic differential equations for the conditional density and the conditional moments were first derived formally in STRATONOVICH [1960] and KUSHNER [1964].] The theorem is analogous to the space-time approach of FEYNMAN [1948] in quantum mechanics, and mathematically, it is an explicit form of the Radon-Nikodym derivative of one function space measure with respect to another (see MORTENSON [1966] and DUNCUN [1967]). The main result in Section 2 generalizes the theorem to a similar role in the general theory of optimal estimation. The proof of this result provided here is simpler and requires weaker hypotheses than any other published proof for the filtering case to the knowledge of the author. This generalization forms a cornerstone for the theory in the sense that it facilitates other results of interest, which are discussed in the remaining sections of the paper.

Thus in Section 3 it is shown that the conditional density of smoothing is equal to the product of the conditional density of filtering and the conditional density of parameter estimation divided by the absolute density and the normalizing constant. A stochastic integral equation is then derived for the conditional density of smoothing. Solving this equation, we obtain the main result of this section, an explicit expression for the conditional density of smoothing in terms of conditional density of filtering.

The first main result of Section 4 is an explicit formula for the conditional density of prediction in terms of the conditional density of filtering. The formula resembles the well-known Chapman-Kolmogorov equation for the transition density. Consequently, this leads to the second main result of this section, a general Chapman-Kolmogorov equation which can easily be proved by the Markov property and the smoothing property of the conditional expectation.

Several major results of FROST [1968] for nonlinear smoothing and prediction through innovation process approach (see equations (3.18) and (4.8)) are recovered as by-products. Some remarks on those results are made, following Corollary (3.3) and Corollary (4.4).

In LO [1969] it is shown that a large class of nonlinear systems can be reduced to a system with linear dynamics and arbitrary a Priori distribution by applying the concept of finite dimensional sensor orbits. The optimal smoothing and prediction of this class of nonlinear systems are resolved in Section 5 with the aid of some of the results in LO [1969]. This section illustrates the applicability of the results of the previous sections in obtaining effectively computable formulas for optimal smoothing and prediction.

In concluding, we remark that all the sequential estimation problems (at least for the models considered in this paper) can be reduced to the problem of filtering. Hence, filtering theory is properly seen to be the core of the estimation theory and, the author believes, the direction in which future research should be focused.

Because of the limited length of the manuscript required by the conference, all proofs are omitted here and will be published elsewhere.

2. A GENERALIZATION OF BUCY REPRESENTATION THEOREM

Throughout this paper all the times considered belong to a finite time interval denoted by S.

(2.1) LEMMA. *Let*

$x.$ *be a real-valued finite dimensional vector Markov process,*

$\pi = \{t_o = \tau_o < \cdots < \tau_n = t\}$ *denote a fixed partition of* $[t_o, t]$,

β_i, $\forall\, i = 1, \ldots, n$, *be a functional on* $\{x_s,\ \tau_{i-1} \le s \le \tau_i\}$,

ν_i, $\forall\, i = 1, \ldots, n$, *be independent, zero-mean, Gaussian and of variance* R_i,

$x.$ *be independent of* ν_i, $\forall\, i = 1, \ldots, n$,

$T =$ *a fixed set* $\{t_1, \ldots, t_m\}$,

$T' =$ *a fixed set* $\{t_1', \ldots, t_{m'}'\}$,

$x(T) = (x_{t_1}, \ldots, x_{t_m})$,

$\eta_T = (\eta_1, \ldots, \eta_m)$,

$\alpha_i = \beta_i + \nu_i$, $\forall\, i = 1, \ldots, n$.

Then the conditional joint probability density

$$p_{x(T)}(\eta_T | \alpha_\pi, x(T') = \xi_{T'}) = \frac{\mathcal{E}_{\mu_x}(\exp H | x(T) = \eta_T, x(T') = \xi_{T'})\, p_{x(T)}(\eta_T | x(T') = \xi_{T'})}{\mathcal{E}_{\mu_x}(\exp H | x(T') = \xi_{T'})} \qquad (2.2)$$

where $H = \sum_{i=1}^{n} \left((\beta_i, \alpha_i)_{R_i^{-1}} - \frac{1}{2} \|\beta_i\|^2_{R_i^{-1}} \right)$ *and* μ_x *is the measure induced on the* sample path space of $\{x_s,\ s \in S\}$ by $\{x_s,\ s \in S\}$.

(2.3) THEOREM. We consider the nonlinear model described by (1.1) and (1.2), where $x.$ is any Markov process of which all its joint probability distribution absolutely continuous with respect to Lebesgue measure. Assume that

$$\mathcal{E}_{\mu_x}\left(\exp\left(2 \int_{t_o}^{t} \|h(x_s, s)\|^2_{R_s^{-1}}\, ds\right)\right) \le M < \infty . \qquad (2.4)$$

Then the joint conditional probability density may be expressed in terms of function space integrals as the following:

$$p_{x(T)}(\eta_T | z_s, t_o \le s \le t, x(T') = \xi_{T'}) = \frac{\mathcal{E}_{\mu_x}(\exp H(t_o,t) | x(T) = \eta_T, x(T') = \xi_{T'}) p_{x(T)}(\eta_T | x(T') = \xi_{T'})}{\mathcal{E}_{\mu_x}(\exp H(t_o,t) | x(T') = \xi_{T'})} , \qquad (2.5)$$

where $H(t_o, t) = -\frac{1}{2} \int_{t_o}^{t} \|h(x_s,s)\|^2_{R_s^{-1}} ds + \int_{t_o}^{t} h'(x_s,s) R_s^{-1} dz_s$ *and the other notations are defined as in Lemma (2.1).*

3. OPTIMAL NONLINEAR SMOOTHING

In this section we consider the nonlinear model described by (1.1) and (1.2) under the hypothesis of Theorem (2.3).

(3.1) THEOREM. *The conditional density of smoothing* $p_{x_\tau}(y | z_s, s \in [t_o, t])$, *for* $t_o \le \tau \le t$, *is equal to the product of the conditional density of filtering* $p_{x_\tau}(y | z_s, s \in [t_o, \tau])$ *and the conditional density of parameter estimation* $p_{x_\tau}(y | z_s, s \in [\tau, t])$ *divided by the density* $p_{x_\tau}(y)$ *and the* normalizing constant K, i.e.,

$$p_{x_\tau}(y | z_s, s \in [t_o, t]) = \frac{p_{x_\tau}(y | z_s, s \in [t_o, \tau]) p_{x_\tau}(y | z_s, s \in [\tau, t])}{K p_{x_\tau}(y)} , \qquad (3.2)$$

where

$$K = \int \frac{p_{x_\tau}(y | z_s, s \in [t_o, \tau]) p_{x_\tau}(y | z_s, s \in [\tau, t])}{p_{x_\tau}(y)} dy .$$

(3.3) LEMMA. *The conditional density of smoothing* $p_{x_\tau}(y | z_s, s \in [t_o, t])$, *for* $t_o \le \tau \le t$, *satisfies the following stochastic integral equation*:

$$p_{x_\tau}(y | z_s, s \in [t_o, t]) = p_{x_\tau}(y | z_s, s \in [t_o, \tau]) + \int_\tau^t p_{x_\tau}(y | z_s, s \in [t_o, \sigma]) \cdot$$

$$\cdot \left(\int_{R^n} h'(\eta, \sigma) p_x(\eta | x_\tau = y, z_s, s \in [t_o, \sigma]) d\eta - \hat{h}'_{\sigma|\sigma}\right) \cdot$$

$$\cdot R_\sigma^{-1} (dz - \hat{h}_{\sigma|\sigma} d\sigma) ,$$

where $\hat{h}_{\sigma|\sigma} = \mathcal{E}(h(x_\sigma, \sigma) | z_s, t_o \le s \le \sigma)$.

(3.4) COROLLARY. *The smoothing estimate* $\hat{x}_{\tau|t}$ *for* $t_o \le \tau \le t$ *satisfies*

$$\hat{x}_{\tau|t} = \hat{x}_{\tau|\tau} + \int_\tau^t \mathcal{E}[x_\tau (h(x_\sigma, \sigma) - \hat{h}_{\sigma|\sigma})' | z_s, s \in [t_o, \sigma])](dz_\sigma - \hat{h}_{\sigma|\sigma} d\sigma). \quad (3.5)$$

REMARK. Equation (3.5) was first derived in FROST [1968] through innovation approach. By definition, the smoothing estimate $\hat{x}_{\tau|t}$ may easily be calculated, given the value of $p_{x_\tau}(y|z_s, s \in [t_o, t])$. Using equation (3.5), the calculation of $\hat{x}_{t|\tau}$ requires the value of $p_{x_\tau}(y|z_s, s \in [t_o, \sigma])$ for all $\sigma \in [\tau, t]$, even if the density of filtering is given. Therefore, the author sees little prospect of the equation for nonlinear smoothing.

(3.6) THEOREM. *The conditional density of smoothing* $p_{x_\tau}(y|z_s, s \in [t_o, t])$, *where* $t_o \leq \tau \leq t$, *may be expressed in terms of the conditional densities of filtering as follows,*

$$p_{x_\tau}(y|z_s, s \in [t_o, t]) = p_{x_\tau}(y|z_s, s \in [t_o, \tau])\exp[\int_\tau^t \alpha_\sigma R_\sigma^{-1}\, dI_\sigma - \frac{1}{2}\int_\tau^t \|\alpha_\sigma\|^2_{R_\sigma^{-1}}\, d\sigma], \tag{3.7}$$

where

$$dI_\sigma = dz_\sigma - \hat{h}_{\sigma|\sigma}\, d\sigma, \tag{3.8}$$

and

$$\alpha_\sigma = \int_{R^n} h'(\eta, \sigma) p_{x_\sigma}(\eta|x_\tau = y, z_s, s \in [\tau, \sigma])d\eta - \hat{h}'_{\sigma|\sigma}. \tag{3.9}$$

REMARK. Equation (3.7) may be derived directly from Theorem (3.1).

4. OPTIMAL NONLINEAR PREDICTION

All through this section we consider the nonlinear model described by (1.1) and (1.2) under the hypothesis of Theorem (2.3), unless otherwise stated.

(4.1) THEOREM. *Given the conditional density of filtering, the conditional density of prediction* $p_{x_\tau}(y|z_s, s \in [t_o, t])$, *for* $t_o \leq t \leq \tau$, *may easily be calculated by the equation*

$$p_{x_\tau}(y|z_s, s \in [t_o, t]) = \int_{R^n} p_{x_t}(\eta|z_s, s \in [t_o, t]) p_{x_\tau}(y|x_t = \eta)d\eta. \tag{4.2}$$

REMARK. Equation (4.2) apparently resembles the well-known Chapman-Kolmogorov equation. (See DOOB [1952].) In fact, these two equations are special cases of the following general theorem.

(4.3) THEOREM. *Let* $x(t)$ *be an arbitrary* n-*vector Markov process (continuous or discrete). Then, given any* σ-*field* Σ_t *which is independent of* $x(s)$ *for all* $s \geq t$, *the conditional density* $p_{x_\tau}(y|\Sigma_t)$ *for* $\tau = t$ *satisfies the equation*

$$p_{x_\tau}(y|\Sigma_t) = \int_{R^n} p_{x_\tau}(y|x_t = \eta)p_{x_t}(\eta|\Sigma_t)d\eta \,. \tag{4.4}$$

(4.5) LEMMA. The conditional density of prediction $p_{x_\tau}(y|z_s, s \in [t_o, t])$, where $t_o \le t \le \tau$, satisfies the following stochastic integral equation,

$$p_{x_\tau}(y|z_s, s \in [t_o, t]) = p_{x_\tau}(y) + \int_{t_o}^{t} [\int_{R^n} h'(\eta, \sigma)p_{x_\tau}(y|x_\sigma = \eta) \cdot$$

$$\cdot\, p_{x_\sigma}(\eta|z_s, s \in [t_o, \sigma])d\eta - p_{x_\tau}(y|z_s, s \in [t_o, \sigma])\hat{h}_{\sigma|\sigma}] \cdot$$

$$\cdot\, R_\sigma^{-1}(dz_\sigma - \hat{h}_{\sigma|\sigma} d\sigma) \,. \tag{4.6}$$

(4.7) COROLLARY. The prediction estimate $\hat{x}_{\tau|t}$ for $t_o \le t \le \tau$ satisfies

$$\hat{x}_{\tau|t} = \mathcal{E}(x_\tau) + \int_{t_o}^{t} \mathcal{E}(x_\tau(h(x_\sigma, \sigma) - \hat{h}_{\sigma|\sigma})' \big| z_s, s \in [t_o, \sigma])(dz_\sigma - \hat{h}_{\sigma|\sigma} d\sigma) \,. \tag{4.8}$$

REMARK. Equation (4.8) was first derived in FROST [1968] through innovation approach. By the reason similar to that in the remark following Corollary (3.4), the author believes that the equation is not of much use in nonlinear prediction.

5. SYSTEMS OF LINEAR DYNAMICS AND ARBITRARY A PRIORI DISTRIBUTION

Let $x(t)$ and $z(t)$ be the n-dimensional signal process and the r-dimensional observation process respectively satisfying the stochastic differential equations

$$dx = F(t)x\,dt + G(t)d\beta_t \,, \tag{5.1}$$

$$x_{t_o} = c \,, \tag{5.2}$$

$$dz = H(t)x\,dt + d\nu_t \,, \tag{5.3}$$

where β_t and ν_t are independent s- and r-dimensional zero mean Brownian motion processes with covariance satisfying

$$\mathcal{E}(\beta_t - \beta_{t_o})(\beta_s - \beta_t)' = \int_{t_o}^{\min(t,s)} Q(\tau)d\tau \tag{5.4}$$

$$\mathcal{E}(\nu_t - \nu_{t_o})(\nu_s - \nu_{t_o})' = \int_{t_o}^{\min(t,s)} R(\tau)d\tau \tag{5.5}$$

and c is an n-dimensional random variable with a density function $g(c)$.

We assume $\int c_i^2 g(c) < \infty$ for all i. The functions $F(t)$, $G(t)$, $H(t)$, $R(t)$, $Q(t)$ are assumed continuous and $\|R_t^{-1}\|$ is assumed uniformly bounded about t_o. To save some hand-waving, we define an operator L as follows:

$$L(f(c_1), t) = \frac{\int f(c_1) g(c_1) \exp(-\frac{1}{2} \|c_1 - V_t M_t\|^2_{V_t^{-1}}) dc_1}{\int g(c_1) \exp(-\frac{1}{2} \|c_1 - V_t M_t\|^2_{V_t^{-1}}) dc_1} .$$

Let us first consider the smoothing problem.

(5.6) THEOREM. *Let* $x(t)$ *and* $z(t)$ *be defined as above. Then the conditional distribution of* $x(\tau)$ *given* z_s, $s \in [t_o, t]$ *for* $t_o \leq \tau \leq t$ *has a density which is*

$$p_{x_\tau}(y | z_s, s \in [t_o, t]) = p_{x_\tau}(y | z_s, s \in [t_o, \tau]) \exp[\int_\tau^t \alpha_\sigma R_\sigma^{-1} dI - \frac{1}{2} \int_\tau^t \|\alpha_\sigma\|^2_{R_\sigma^{-1}} d\sigma] , \tag{5.7}$$

where

$$p_{x_\tau}(y | z_s, s \in [t_o, \tau]) = |\det p_o^{-1}|^{\frac{1}{2}} L(\exp - \frac{1}{2} \|y - \Phi(\tau, t_o) c_1 - \hat{\hat{x}}_o(t)\|^2_{p_o^{-1}}, \tau) \tag{5.8}$$

$$\alpha_\sigma = [H_\sigma(\hat{x}_{\tau o}(\sigma) - \hat{x}_\sigma + \Phi(\sigma, \tau) y)]' , \tag{5.9}$$

$$dI_\sigma = dz_\sigma - H_\sigma \hat{x}_\sigma d\sigma , \tag{5.10}$$

$$\hat{x}_t = \hat{x}_o(t) + (\Phi(t, t_o) + A_t) L(c_1, t) , \tag{5.11}$$

$$d\hat{x}_{\tau o}(t) = F(t) \hat{x}_{\tau o}(t) dt + P_{\tau o}(t) H(t) R_t^{-1} (dz(t) - H(t) \hat{x}_{\tau o}(t)) , \tag{5.12}$$

$$\hat{x}_{\tau o}(\tau) = y , \tag{5.13}$$

$$\dot{P}_{\tau o}(t) = F P_{\tau o}(t) + P_{\tau o}(t) F' - P_{\tau o}(t) H' R^{-1} H P_{\tau o}(t) + G Q G' , \tag{5.14}$$

$$P_{\tau o}(\tau) = 0 , \tag{5.15}$$

$$\hat{\hat{x}}_o(t) = A_t c_1 + \hat{x}_o(t) , \tag{5.16}$$

$$A_t = - \int_{t_o}^t \psi(t, s) P_o(s) H'(s) R_s^{-1} H(s) \Phi(s, t_o) ds , \tag{5.17}$$

$$d\hat{x}_o = F(t) \hat{x}_o dt + P_o H' R^{-1} (dz - H \hat{x}_o dt) , \quad \hat{x}_o(t_o) = 0 , \tag{5.18}$$

$$\dot{P}_o = FP_o + P_oF' - P_oH'R^{-1}HP_o + GQG', \qquad P_o(t_o) = 0, \tag{5.19}$$

$$V_t^{-1} = \int_{t_o}^{t} (A_s + \Phi(s, t_o))'H_s'R_s^{-1}H_s(A_s + \Phi(s, t_o))ds, \tag{5.20}$$

$$M_t = \int_{t_o}^{t} (A_s + \Phi(s, t_o))'H_s'R_s^{-1}(dz_s - H_s\hat{x}_{os}\,ds), \tag{5.21}$$

$\Phi(t, s)$ *is the transition matrix of* F_t *and* $\psi(t, s)$ *is the transition* matrix of $F_t - P_{ot}H_t'R_t^{-1}H_t$.

(5.22) THEOREM. *Let all symbols be defined as in the last theorem. Then the conditional smoothing estimate and its conditional error covariance satisfy*

$$\hat{x}_{\tau|t} = D(\tau, t)P_{o\tau}^{-1}\hat{x}_{o\tau} + D(\tau, t)[\int_{\tau}^{t} R_\sigma^{-1}H_\sigma\Phi(\sigma, \tau)dz_\sigma +$$
$$-\frac{1}{2}\int_{\tau}^{t} \Phi\,(\sigma, \tau)H_\sigma R_\sigma^{-1}H_\sigma(z\hat{x}_{\tau o}(\sigma) - \hat{x}_\sigma)d\sigma] + D(\tau, t)P_{o\tau}^{-1}(\Phi(\tau, t_o) + A_\tau)L_1(c_1). \tag{5.23}$$

$$\mathcal{E}[(x_\tau - \hat{x}_{\tau|t})(x_\tau - \hat{x}_{\tau|t})'|z_s, s \in [t_o, t]]$$
$$= D_{(\tau,t)} + D_{(\tau,t)}P_{o\tau}^{-1}(\Phi(\tau, t_o) + A_\tau)L_1(c_1c_1')(\Phi(\tau, \tau_o) + A_\tau)'P_{o\tau}^{-1}D_{(\tau,t)} +$$
$$- D_{(\tau,t)}P_{o\tau}^{-1}(\Phi(\tau, t_o) + A_\tau)L_1(c_1)L_1(c_1')(\Phi(\tau, t_o) + A_\tau)'P_{o\tau}^{-1}D(\tau, t), \tag{5.24}$$

where

$$L_1(f(c_1)) = \frac{\int f(c_1)g(c_1)\gamma(c_1)dc_1}{\int g(c_1)\gamma(c_1)dc_1}, \tag{5.25}$$

$$D_{(\tau,t)}^{-1} = P_{o\tau}^{-1} + \int_{\tau}^{t} \Phi'(\sigma, \tau)H_\sigma'R_\sigma^{-1}H_\sigma\Phi(\sigma, \tau)d\sigma, \tag{5.26}$$

$$\gamma(c_1) = \exp\{-\frac{1}{2}\|c_1 - V_\tau M_\tau\|^2_{V_\tau^{-1}} - \frac{1}{2}\|(\Phi(\tau, t_o) + A_\tau)c_1 + \hat{x}_{o\tau}\|^2_{P_{o\tau}^{-1}} +$$
$$+\frac{1}{2}\|P_{o\tau}^{-1}[(\Phi(\tau, t_o) + A_\tau)c_1 + \hat{x}_{o\tau}]\|^2_{D(\tau,t)}\}. \tag{5.27}$$

REMARK. The smoothing problem of the Kalman-Bucy model is a special case of the above results.

(5.28) THEOREM. *Let all the symbols be defined as in Theorem (5.6). Then the conditional predicting estimate and its conditional error covariance, for* $t \leq \tau$,

$$\hat{x}_{\tau|t} = \Phi(\tau, t)\hat{x}_o(t) + (\Phi(\tau, t_o) + \Phi(\tau, t)A_t)L(c_1, t) , \qquad (5.29)$$

$$\begin{aligned} \mathcal{E}[(x_\tau - \hat{x}_{\tau|t})(x_\tau - \hat{x}_{\tau|t})'|z_s, s \in [t_o, t]] &= C_Q(t, \tau) + \Phi(\tau, t)P_{ot}\, \Phi'(\tau, t) + \\ &+ (\Phi(\tau, t_o) + \Phi(\tau, t)A_t)\; L(c_1 c_1', t) - [L(c_1, t)][L(c_1', t)] \cdot \\ &\cdot (\Phi(\tau, t_o) + \Phi(\tau, t)A_t)' , \end{aligned} \qquad (5.30)$$

where

$$C_Q(t, \tau) = \int_t^\tau \Phi(\tau, \sigma)G(\sigma)Q(\sigma)G'(\sigma)\Phi'(\tau, \sigma)d\sigma . \qquad (5.31)$$

This paper was supported in part by AFOSR Grant 70-1898 and Army Contract DA-31-124-ARO(D)394, and NASA Grant NGL-05-020-073.

REFERENCES

BUCY, R. S.

[1970] Linear and nonlinear filtering, to appear.

BUCY, R. S. and JOSEPH, P. D.

[1968] Filtering for Stochastic Processes with Application to Guidance, Interscience Publishers, A Division of John Wiley and Sons, New York.

BUCY, R. S.

[1965] Nonlinear filtering theory, IEEE Trans., PGAC, AC-10:198.

DUNCAN, C. E.

[1967] Probability densities for diffusion processes, AFOSR Tech. Report 7001-4, Stanford University.

FEYNMAN, R. P.

[1948] Space-time approach to non-relativistic quantum mechanics, Review of Modern Physics, 20:267.

FROST, P. A.

[1968] Estimation in continuous-time non-linear systems, Thesis, Department of Electrical Engineering, Stanford University.

ITO, K.

[1951] On stochastic differential equations, Mem.-Amer. Math. Soc., 4.

KALMAN, R. E. and BUCY, R. S.

[1961] New results in linear filtering and prediction theory, J. Basic Engineering, 83:95-108.

KALMAN, R. E.

[1960] A new approach to linear filtering and prediction problems, J. Basic Engineering, 82:34-45.

KUSHNER, H. J.

[1962] On the differential equations satisfied by conditional probability densities of Markov processes, with applications, SIAM J. on Control, 2:106-119.

LO, J. T.

[1969] Finite dimensional sensor orbits and optimal nonlinear filtering, Thesis, Department of Aerospace Engineering, University of Southern California.

LOÉVE, M.

[1963] Probability Theory, 3rd. Ed., Van Nostrand Co., Princeton.

MORTENSEN, R. E.

[1966] Optimal control of continuous-time stochastic systems, Thesis, Report no. ERL-66-1, Electronics Research Laboratory, University of California, Berkeley.

STRATONOVICH, R. L.

[1968] Conditional Markov Processes and Their Application to the Theory of Optimal Control, Translated from the Russian by R. N. and N. B. McDonough for Scripta Technica, Inc.

[1960] Conditional Markov processes theory, Prob. Appl., 5:156-178.

SKOROKHOD, A. V.

[1961] Studies in the Theory of Random Processes. Addison Wesley, Reading, Mass.

SYSTEM IDENTIFICATION BY QUASILINEARIZATION USING WEIGHTING COEFFICIENTS

C. K. ALEXANDER
G.V.S. RAJU
Department of Electrical Engineering
Ohio University
Athens, Ohio 45701

ABSTRACT

This paper is concerned with the determination of an optimal set of weighting coefficients to be used in the design of an identifier based on quasilinearization. The weighting coefficients are optimal in the sense that the identification process converges in a minimum number of iterations. Examples are given to illustrate the advantages of using the optimal weighting coefficients in the quasilinearization method.

INTRODUCTION

In order to exercise the proper control of dynamic systems, it is first necessary to identify the plant dynamics. Conventional identification techniques used in identifying system transfer function or impulse response are limited to linear, time invariant systems. The quasilinearization method identifies the system in terms of its describing differential equation. This approach is not only good for linear and time invariant systems but also for nonlinear, time variant systems. Kumar and Sridhar [1] and Eveleigh [2] have used quasilinearization methods for identification of systems. Kumar and Sridhar apply quasilinearization directly. Their method may not converge to the proper values if the initial guessed parameters are not close enough to the actual parameters and/or the observation times are not correct.

Eveleigh uses a performance index of the type

$$P = \int_{t_o}^{t_o + T} \sum_{i=1}^{N} g_i \{x_i(x^\circ) + \sum_{j=1}^{N} \delta x_j^\circ \phi_{ij} - \hat{x}_i\}^2 dt.$$

In the above performance index T is the identification interval, g_i's are the weighting coefficients, $\hat{x}_i$ is the observed system response, $x_i(x^\circ)$ is the corresponding model response to the initial conditions, x°, and $\delta x_j^\circ \phi_{ij}$ is the linearized perturbation or approximation of the difference between $x_i(x^\circ)$ and $\hat{x}_i$. In the above method T and g_i are chosen arbitrarily.

With arbitrarily chosen g_i's the method may not always lead to proper identification even after an inordinate number of iterations. In fact, for some cases the identification process may even diverge.

This paper is concerned with the determination of an optimal set of g_i's for a sampling interval T* so that the identification process converges in a minimum number of iterations. This improved method allows larger deviations between the actual values and the guessed values than Kumar and Sridhar's method. Examples are given to illustrate the advantages of using the optimal weighting coefficients in the quasilinearization method.

IDENTIFICATION BY QUASILINEARIZATION

The problem is to choose model parameters and initial states to minimize the measured difference between system and model responses. The resulting parameters provide the desired system identification.

The assumed model is of the form

$$\dot{x} = f(a,x), \quad x(t_o) = x^\circ \tag{1}$$

where

a is a constant L x 1 parameter vector

$x = N_o$ x 1 state vector.

Since values of x° and a are chosen to minimize the measured difference, it is convenient to treat x and a alike. For this purpose we define the auxiliary relations

$$\dot{a} = 0$$

and henceforth treat a as part of the state vector. Thus we have

$$\dot{x} = f(x), \quad x(t_o) = x^\circ \tag{2}$$

where

x = N x 1 state vector

and

$N = L + N_o$.

We wish to minimize the following performance index

$$P = \int_{t_o}^{t_o+T} \sum_{i=1}^{N} g_i \{x_i(x^\circ) - \hat{x}_i(\hat{x}^\circ)\}^2 dt \tag{3}$$

by choice of initial state and parameter vectors. The g_i's are weighting coefficients which are equal to zero for unmeasured elements of the state. $\hat{x}_i(\hat{x}^\circ)$'s are the actual observed response of the system and $\hat{x}^\circ$ is the actual system initial conditions (an N x 1 vector).

Let δx° denote a nominal change in the initial conditions. Then the model response to the initial conditions $x^\circ + \delta x^\circ$ is given by

---

*T may be chosen arbitrarily or it may be chosen optimally by minimizing P(G) with respect to T.

$$x_i(x^\circ + \delta x^\circ) \cong x_i(x^\circ) + \sum_{j=1}^{N} \delta x_j^\circ \phi_{ij}$$

$i = 1, 2, ---, N.$

where ϕ is the fundamental matrix obtained by solving the matrix equation $\dot{\phi}=F\phi$. The ij^th^ element of F is given by $df_i(x)/dxj$ evaluated along the nominal solution path [2]. Then equation (3) becomes

$$P = \int_{t_o}^{t_o+T} \sum_{i=1}^{N} g_i\{x_i(x^\circ) + \sum_{j=1}^{N} \delta x^\circ \phi_{ij} - \hat{x}_i(\hat{x}^\circ)\}^2 dt \quad . \tag{4}$$

It is necessary for minimum P with respect to δx_j° that

$$\frac{\partial P}{\partial \delta x_j^\circ} = 0 = 2\int_{t_o}^{t_o+T} \sum_{i=1}^{N} g_i\{(x_i(x^\circ) + \sum_{k=1}^{N} \delta x_k^\circ \phi_{ik} - \hat{x}_i(\hat{x}^\circ))\phi_{ij}\}dt \tag{5}$$

for j = 1,2, ---, N. In Eveleigh's method the g_i's are assumed. Therefore, all functions in the integrand of equation (5) are known except for the δx_k°'s; equation (5) provides N algebraic relationships defining the desired δx_k°. This iterative process is repeated, based on the revised initial conditions (or states)

$$x^\circ_{new} = x^\circ_{old} + \delta x^\circ, \tag{6}$$

until improvements become negligible.

METHOD OF DETERMINING THE WEIGHTING COEFFICIENTS

As stated earlier the objective of this paper is to find an optimal* set of g_i, so that the identification process converges in a minimum number of iterations. In other words we are interested in the determination of g_i to be used in the design of an identifier based on Eveleigh's method. This means that in equation (5) the g_i's are unknown (unlike Eveleigh's method) and hence the δx° is a function of g_i also. From equation (5) we get

$$\int_{t_o}^{t_o+T} \sum_{i=1}^{N} g_i(\hat{x}_i(\hat{x}^\circ) - x_i(x^\circ))\phi_{ij} dt$$

---

* Optimum in the sense that the performance index (equation 14) is minimized.

$$= \int_{t_o}^{t_o+T} \sum_{i=1}^{N} g_i \left(\sum_{k=1}^{N} \delta x_k^\circ \phi_{ik} \right) \phi_{ij}\, dt. \tag{7}$$

Let

$$A_{jk} = \sum_{i=1}^{N} \int_{t_o}^{t_o+T} g_i \phi_{ik} \phi_{ij}\, dt \ , \tag{8}$$

$$C_j = \sum_{i=1}^{N} \int_{t_o}^{t_o+T} g_i \{\hat{x}_i(\hat{x}^\circ) - x_i(x^\circ)\}\ \phi_{ij}\, dt \ , \tag{9}$$

and

$$G = [g_1,\ g_2,\ ---,\ g_{No},\ 0,\ ---,\ 0].$$

Note that because of the nature of equation (5) G can be normalized with respect to g_1, therefore

$$G = [1, g_2,\ ---,\ g_{No},\ 0,\ ---,\ 0]. \tag{10}$$

Then equation (7) becomes

$$C(G,\ x^\circ,\ \hat{x}^\circ) = A(G,\ x^\circ)\ \delta x^\circ \tag{11}$$

or

$$\delta x^\circ = A^{-1} C = \delta x^\circ (G,\ x^\circ,\ \hat{x}^\circ). \tag{12}$$

The K+1 th revised initial condition is obtained from the Kth via

$$[x^\circ]^{K+1} = [x^\circ]^{K} + \delta x^\circ (G,\ [x^\circ]^{K},\ \hat{x}^\circ). \tag{13}$$

Now the problem is to find the optimum g_i for rapid convergence of the identification process for all possible initial (x°) and actual ($\hat{x}^\circ$) states.

The quasilinearization method gives best results when the perturbations between the actual values and the initial values are small. One can determine from the initial design the possible limits of these perturbations. The convergence of the identification process may become worse as the initial state (x°) approaches one limit and the actual system value ($\hat{x}^\circ$) approaches the other limit. Thus if one can find values of g_i's that are optimum with respect to these extremes then it is reasonable to assume that these g_i's will be optimum or near optimum for all other intermediate values. The method presented here is to determine a set of g_i's which will yield convergence of the identification process for the extreme cases. This will then lead to the design of an identifier to be used on the unknown system. To obtain an optimum set of g_i's, a least squares fit performance index of the following type is chosen.

$$\text{Performance index } P(G) = \sum_{i=1}^{M} ([x^\circ]^K - \hat{x}^\circ)_i^2 \qquad (14)$$

where $M = 2^N$ since x° and $\hat{x}^\circ$ are $N \times 1$ vectors

i = the index of the i th extreme case of x° and $\hat{x}^\circ$

K = the number of iterations before $P(G)$ is evaluated.

The performance index is a measure of the squares of differences between initial and actual states after K iterations. The value of K should be chosen large enough to allow the identification process to converge for some value of G. Minimizing (14) with respect to g_i will result in an optimal set of g_i. The minimization of (14) can be accomplished by using any of the existing gradient techniques. The method of determining the optimal weighting coefficients is illustrated by the following examples.

EXAMPLES

Example 1

The purpose of this linear example is to illustrate the method.

$$\dot{x}_1 = ax_1 - x_2$$

$$\dot{x}_2 = - x_2$$

Let $a = x_3$. Then the model equations become

$$\dot{x}_1 = x_3x_1 - x_2, \quad x(t_o) = x^\circ.$$

$$\dot{x}_2 = -x_2$$

$$\dot{x}_3 = 0$$

Let the range of x° and $\hat{x}^\circ$ be

$$1 \leq (x_1^\circ, \hat{x}_1^\circ) \leq 3$$

$$1 \leq (x_2^\circ, \hat{x}_2^\circ) \leq 3$$

$$-.5\text{-DA} \leq (x_3^\circ, \hat{x}_3^\circ) \leq -.5\text{+DA}$$

The value of DA was varied in order to observe the effect of increasing the parameter space on P(G). In this case there are 8 ($M = 2^3 = 8$) extreme cases of x° and $\hat{x}^\circ$. These are shown in the following table.

Case	x_1°	$\hat{x}_1^\circ$	x_2°	$\hat{x}_2^\circ$	x_3°	$\hat{x}_3^\circ$
1	1	3	1	3	-.5+DA	-.5-DA
2	1	3	1	3	-.5-DA	-.5+DA
3	1	3	3	1	-.5+DA	-.5-DA
4	1	3	3	1	-.5-DA	-.5+DA
5	3	1	1	3	-.5+DA	-.5-DA
6	3	1	1	3	-.5-DA	-.5+DA
7	3	1	3	1	-.5+DA	-.5-DA
8	3	1	3	1	-.5-DA	-.5+DA

Using K=3 P(G) is formed from equation (14) as

$$P(G) = \sum_{i=1}^{8} \{[x^\circ]^3 - \hat{x}^\circ\}_i^2$$

G optimal is found by using the Newton-Raphson one dimensional minimization technique.

$$G \text{ optimal} = [1,\ 10^{-3},\ 0]$$

For illustrative purposes plots of P(G) vs g_2 and P(G) vs log (g_2) are shown in figures one and two. In this case assignment of a weight to the second component is important. Even in a linear system of this type there is an optimum set of g_i's which yield better identification.

Example 2

The purpose of this example is to illustrate the importance of the weighting coefficients, g_i, on a simple nonlinear problem.

$$\dot{x}_1 = a\ , \quad x(0) = x^\circ.$$

$$\dot{x}_2 = x_1 x_2$$

The limits were taken as

$$-1 \le (x_1^\circ,\ \hat{x}_1^\circ) \le 1$$

$$-10 \le (x_2^\circ,\ \hat{x}_2^\circ) \le 10$$

$$-DA \le (x_3^\circ,\ \hat{x}_3^\circ) \le DA$$

Using K=3 P(G) is formed. G optimal is found by using the one dimensional Newton-Raphson minimization technique.

$$G \text{ optimal} = [1, g_2 << 1,\ 0]$$

Table 1 shows P(G) vs g_2 for various values of DA.

In reference [3] it is shown analytically that for $g_2 << 1$ x_1° and x_3° converge in one iteration and x_2° converges in at most two iterations.

Example 3 - Van der Pol Equation

The purpose of this example is to compare the method developed in this paper and the method used by Kumar and Shridhar. Let the augmented Van der Pol equation with $x_3 = \varepsilon$ and $x_4 = \mu$ be [1]

$$\overset{\circ}{x}_1 = x_2$$

$$\overset{\circ}{x}_2 = -x_3(1-x_1^2)\, x_2 - x_4\, x_1$$

$$\overset{\circ}{x}_3 = 0$$

$$\overset{\circ}{x}_4 = 0$$

Consider the set of values that Kumar and Shridhar used in their Example [1].

i.e. $x^\circ = [1.9, -.15, 2, .5]^T$ and $\hat{x}^\circ = [2, 0, 3.65, 1.17]^T$

Applying the improved method to the above case (using Newton-Raphson to minimize $P(G)$ with respect to G and T) resulted in $G = [1, .618, 0]$ and $T = .42$ sec. which in turn yielded convergence in three iterations. This is an improvement over the six iterations required by the method used by Kumar and Shridhar.

Some very interesting things happen when the initial guessed value of x_1° is changed to 1.8 from its original value of 1.9. Kumar and Shridhar's method was then applied with only this one change and after six iterations converges to a false set of values $x^\circ = [2.000, 0.000, -1.417, 9.156]$. These values do indeed satisfy the boundry conditions at the sampling times used (see Figure 3 and reference [3]).

Eveleigh's method (with $T=.42$ sec. and $G = [1, .618, 0]$) was applied to the modified set of values ($x_1^\circ = 1.8$) and it converged in five iterations.

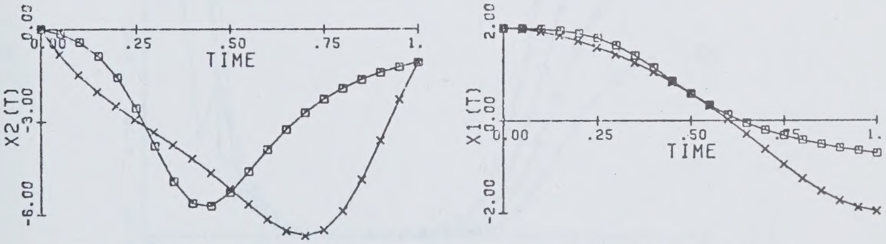


Figure 3. Response of the Van der Pol Equation with the actual parameters (□) and with the false parameters (×).

CONCLUSION

A method has been presented for the determination of an optimal set of weighting coefficients to be used in the design of an identifier based on quasilinearization. G optimal was found by minimizing a performance index which is a measure of the squares of the differences between initial and actual states after K iterations. Examples included in this paper illustrate the advantages of using this improved method, for example, faster convergence. With the improved method it was found that the initial guessed values can deviate farther from the actual values and still converge. Since P(G) can be made a function of the sampling period (T), an optimum T can be found by minimizing P(G) with respect to T. This results in the added benefit that T can also be determined prior to design, unlike other methods where it is chosen arbitrarily.

REFERENCES

[1] Kumar, K.S.P., and R. Sridhar: "On The Identification of Control Systems by the Quasilinearization Method", IEEE Trans. Autom. Control, April, 1964.

[2] Eveleigh, Virgil W., Adaptive Control and Optimization Techniques, McGraw-Hill, 1967, pp. 163-171.

[3] Alexander, Jr., Charles K., "System Identification", Ph.D. Electrical Engineering dissertation, Ohio University, (in preparation).

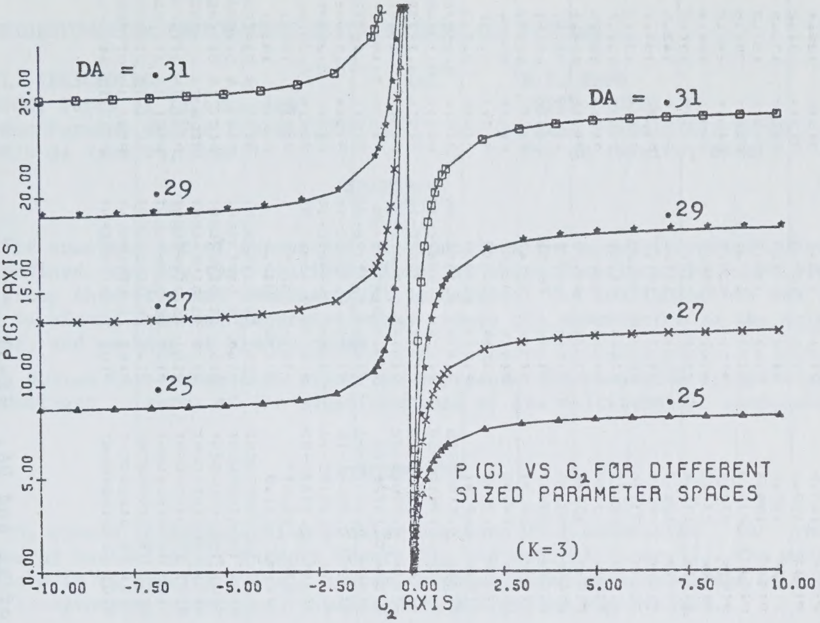


Figure 1.

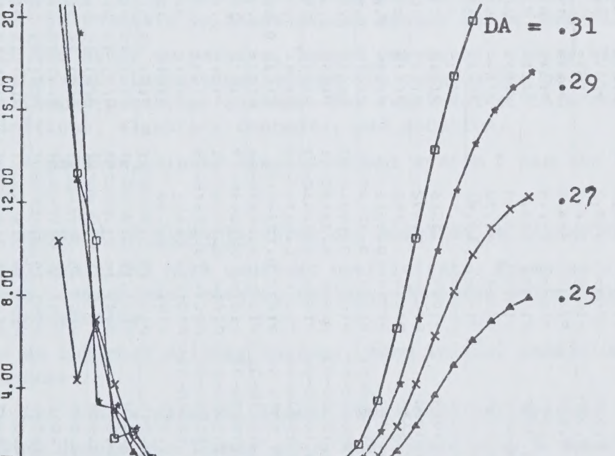


Figure 2. P(G) vs Log g_2 for various values of DA. (K=3)

```
DA=     0.5000000000                                DA=      1.000          DA=      1.500          DA=      2.000          DA=      2.500
G1= 1 AND G2 =    10.00      P(G)=    695.9         P(G)=    467.0          P(G)=    163.6          P(G)=    115.0          P(G)=    134.4
G1= 1 AND G2 =    9.000      P(G)=    671.2         P(G)=    459.2          P(G)=    158.5          P(G)=    107.5          P(G)=    131.2
G1= 1 AND G2 =    8.000      P(G)=    661.5         P(G)=    461.8          P(G)=    168.0          P(G)=    111.9          P(G)=    132.0
G1= 1 AND G2 =    7.000      P(G)=    663.4         P(G)=    473.0          P(G)=    194.4          P(G)=    137.0          P(G)=    151.2
G1= 1 AND G2 =    6.000      P(G)=    670.4         P(G)=    487.0          P(G)=    235.3          P(G)=    187.1          P(G)=    201.5
G1= 1 AND G2 =    5.000      P(G)=    669.8         P(G)=    489.2          P(G)=    273.2          P(G)=    238.8          P(G)=    248.2
G1= 1 AND G2 =    4.000      P(G)=    640.3         P(G)=    453.1          P(G)=    270.2          P(G)=    235.7          P(G)=    224.4
G1= 1 AND G2 =    3.000      P(G)=    550.9         P(G)=    351.1          P(G)=    203.6          P(G)=    172.1          P(G)=    167.8
G1= 1 AND G2 =    2.000      P(G)=    373.8         P(G)=    198.1          P(G)=    121.3          P(G)=    121.3          P(G)=    143.5
G1= 1 AND G2 =    1.000      P(G)=    131.6         P(G)=    81.46          P(G)=    77.19          P(G)=    101.1          P(G)=    137.4
G1= 1 AND G2 =   0.9000      P(G)=    108.5         P(G)=    73.87          P(G)=    73.30          P(G)=    98.20          P(G)=    135.3
G1= 1 AND G2 =   0.8000      P(G)=    86.93         P(G)=    66.30          P(G)=    68.73          P(G)=    94.09          P(G)=    131.7
G1= 1 AND G2 =   0.7000      P(G)=    67.22         P(G)=    58.21          P(G)=    62.99          P(G)=    83.01          P(G)=    125.5
G1= 1 AND G2 =   0.6000      P(G)=    49.82         P(G)=    48.97          P(G)=    55.44          P(G)=    78.70          P(G)=    114.5
G1= 1 AND G2 =   0.5000      P(G)=    34.55         P(G)=    38.00          P(G)=    45.45          P(G)=    64.67          P(G)=    95.41
G1= 1 AND G2 =   0.4000      P(G)=    20.16         P(G)=    24.90          P(G)=    33.17          P(G)=    46.21          P(G)=    66.51
G1= 1 AND G2 =   0.3000      P(G)=    8.892         P(G)=    10.87          P(G)=    20.27          P(G)=    32.47          P(G)=    47.52
G1= 1 AND G2 =   0.2000      P(G)=    8.708         P(G)=    17.68          P(G)=    24.95          P(G)=    35.82          P(G)=    69.94
G1= 1 AND G2 =   0.1000      P(G)=  0.2515          P(G)=  0.2184           P(G)=  0.3192           P(G)=    1.012          P(G)=    3.892
G1= 1 AND G2 =   0.1000D-01  P(G)=  0.3068E-01      P(G)=  0.8959E-01       P(G)=  0.2741           P(G)=  0.8059           P(G)=    2.223
G1= 1 AND G2 =   0.1000D-02  P(G)=  0.1031E-05      P(G)=  0.2962E-05       P(G)=  0.7109E-05       P(G)=  0.1458E-04       P(G)=  0.2652E-04
G1= 1 AND G2 =   0.1000D-03  P(G)=  0.2313E-11      P(G)=  0.3170E-11       P(G)=  0.2410E-10       P(G)=  0.6082E-10       P(G)=  0.1356E-09
G1= 1 AND G2 =   0.1000D-04  P(G)=  0.2543E-17      P(G)=  0.9268E-17       P(G)=  0.2338E-16       P(G)=  0.7502E-16       P(G)=  0.1773E-15
G1= 1 AND G2 =   0.1000D-05  P(G)=  0.2575E-23      P(G)=  0.9391E-23       P(G)=  0.2890E-22       P(G)=  0.7667E-22       P(G)=  0.1824E-21
G1= 1 AND G2 IS << THAN G1   P(G)= 0.1924E-29       P(G)= 0.6804E-29        P(G)= 0.1972E-29        P(G)= 0.4930E-30        P(G)= 0.4930E-29
G1= 1 AND G2 = -0.1000D-05   P(G)=  0.2577E-23      P(G)=  0.9396E-23       P(G)=  0.2897E-22       P(G)=  0.7717E-22       P(G)=  0.1837E-21
G1= 1 AND G2 = -0.1000D-04   P(G)=  0.2598E-17      P(G)=  0.9539E-17       P(G)=  0.2948E-16       P(G)=  0.7880E-16       P(G)=  0.1890E-15
G1= 1 AND G2 = -0.1000D-03   P(G)=  0.2868E-11      P(G)=  0.1090E-10       P(G)=  0.3519E-10       P(G)=  0.9962E-10       P(G)=  0.2582E-09
G1= 1 AND G2 = -0.1000D-02   P(G)=  0.9787E-05      P(G)=  0.7006E-04       P(G)=  0.7047E-03       P(G)=  0.4168E-01       P(G)=  0.3926E-01
G1= 1 AND G2 = -0.1000D-01   P(G)=    1.263         P(G)=    319.4          P(G)=    70.81          P(G)=    42.09          P(G)=    1997.
G1= 1 AND G2 =   -1.000      P(G)=    59.60         P(G)=    1271.          P(G)=    231.0          P(G)=    312.9          P(G)=    1426.
G1= 1 AND  G2=   -2.000      P(G) IS DIVERGENT      P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT
G1= 1 AND  G2=   -3.000      P(G) IS DIVERGENT      P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT
G1= 1 AND  G2=   -4.000      P(G) IS DIVERGENT      P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT
G1= 1 AND  G2=   -5.000      P(G) IS DIVERGENT      P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT
G1= 1 AND  G2=   -6.000      P(G) IS DIVERGENT      P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT
G1= 1 AND  G2=   -7.000      P(G) IS DIVERGENT      P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT
G1= 1 AND  G2=   -8.000      P(G) IS DIVERGENT      P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT
G1= 1 AND  G2=   -9.000      P(G) IS DIVERGENT      P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT
G1= 1 AND  G2=   -10.00      P(G) IS DIVERGENT      P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT       P(G) IS DIVERGENT
```

Table 1. P(G) for various values of g_2 and DA. (K=3)

EIGENFUNCTION PROPOGATION IN INTERCONNECTED SYSTEMS

L. KERSCHBERG
PUC, Dept. of Informatica
Rua Marques de São Vicente,209
Rio de Janeiro, Brasil

and

W.W. HAPP
COPPE - UFRJ
Caixa Postal 1191-ZC-00
Rio de Janeiro, Brasil

ABSTRACT

The complete set of characteristic functions for a multiterminal system is defined, and the test configurations for these functions are enumerated by group theoretic and combinatorial techniques. The configurations are coded in binary form and generator masks reduce the enumeration to the union, shift, and masking of binary codes.

A system interconnection algorithm expresses the composite system's eigenfunctions in terms of the eigenfunctions of its multiterminal components.

I. INTRODUCTION

The system interconnection problem has been well documented in the areas of Mechanics [1], Network Theory[2], and Control Theory[3]. The main concern is to predict overall system response based upon knowledge of mechanical receptances, electrical admittances, and system eigenvalues.

In this paper we present a computer-oriented, unified approach to the system interconnection problem. We show that the set of characteristic polynomials of a system is sufficient to construct admittance matrices, transfer-function matrices, and eigenfunctions for multiterminal systems. Moreover, the polynomials for a composite system may be determined from the component systems' polynomials by means of the system interconnection algorithm.

We restrict our study to passive, lumped parameter, linear time-invariant systems, but there are indications[4]that the results can be extended to active and distributed-parameter systems. The remainder of this section is devoted to definitions, algebraic concepts, and notation.

Definition 1 A passive, linear time-invariant system S has the following properties:

(i) S is composed of elements which are modelled by linear differential (of difference) equations with constant coefficients. Examples are resistors, inductors, capacitors, masses, springs, dampers, potentiometers, integrators, and delay units.

(ii) S has no internal driving sources, zero initial conditions, and no mutual inductances.

Definition 2 Let S be a passive, linear time-invariant system.

(i) Let G(S) denote the linear graph associated with S. Each element of S defines a branch of G(S), and each branch is connected to two nodes of G(S).

(ii) A terminal of S is an access point to a system node. The terminal may be used to excite the system and/or measure system response.

(iii) a port variable exists between every two terminals of S and has value "1" when its terminals are connected together (short-circuited) and value "0" otherwise.

Notation Henceforth, when we speak of an n-terminal system S, we mean a passive, linear time-invariant system which has n terminals.

Fact For $n \geq 2$, an n-terminal system has $n(n-1)/2$ port variables.

Definition 3 Let S be an n-terminal system. A (test) configuration for S is a specified "short-circuit" configuration of its terminals. Each configuration has a corresponding characteristic polynomial which can be calculated by topological methods[5].

Let C(S) denote the set of configurations of S, while C(n) denotes the number of configurations for an n-terminal system.

Fact Let S be an n-terminal system. The number of configurations for **S** can be calculated recursively by the equation

$$C(n+1) = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix} C(n-k) \ , \ C(0)=1 \ ; \begin{bmatrix} n \\ k \end{bmatrix} = \frac{n!}{(n-k)! \, .k!} \ ; \ 0! = 1 \qquad (1)$$

Another representation is $C(n)=(1+x)^{n-1}$ with $x^{k} = C(k)$ and $C(1)= 1$. The chart below relates n and C(n):

n	1	2	3	4	5	6
C(n)	1	2	5	15	52	203

(2)

Definition 4 Let X be a set. An equivalence relation $E \subsetneq X$ x X satisfies the following properties:

(i) Reflexive: xEx $\forall\, x \in X$
(ii) Symmetric: $xEy \rightarrow yEx$ $\forall\, x,y \in X$
(iii) Transitive: xEy & $yEz \rightarrow xEz$ $\forall\, x,y,z \in X$

The relation E partitions X into equivalence classes of the form

$$C(x) = \{ y \mid yEx, \ y \ \varepsilon X\}.$$

Definition 5 A partial ordering of X is a relation $\lesssim \subseteq X$ x X which is reflexive, transitive, and anti-symmetric: $x \lesssim y \not\rightarrow y \lesssim x, \forall x,y \ \varepsilon X$.

Definition 6 Let n be a natural number. The partition of n, *P*(n), is a set of sequences of natural numbers, each sequence arranged in monotonically descending order, with each sequence summing to n.

For example, *P*(5) = {2,2,1; 2,1,1,1; 1,1,1,1,1; 3,1,1; 3,2; 4,1; 5}.

Fact The partition of n, *P*(n), can be partially ordered by the relation

$\lesssim \subseteq P(n)$ x $P(n)$ such that for any p_1, p_2 $\varepsilon P(n)$,

$p_1 \lesssim p_2$ if and only if (i) length $(p_1) \geq$ length (p_2)
and/or
(ii) first digit $(p_1) \lesssim$ first digit (p_2)

Thus, *P*(5) in partially ordered form is

P(5) = {1,1,1,1,1; 2,1,1,1; 2,2,1; 3,1,1; 3,2; 4,1; 5} (3)

Definition 7 (a) A cyclic group of finite order is a group G whose elements are generated by a single element g such that

$$G = \{g^0, g^1, g^2, \ldots, g^{n-1}\}; \quad g^n = g^0 = 1 \text{, with n the order of G.} \tag{4}$$

(b) A group G acts on a set X by means of the function $G \times X \to X$ written $(g,x) \to gx$ such that $1x = x$ and $(g_1 g_2)x = g_1(g_2 x)$. (5)

Fact Let $\underline{n} = \{1,2, \ldots, n\}$ and define the shift operator α: $\underline{n} \to \underline{n}$ as the permutation

$$\alpha = \begin{bmatrix} 1 & 2 & 3 & \ldots & n \\ 2 & 3 & 4 & \ldots & 1 \end{bmatrix} \text{ with } \alpha^o = 1_{\underline{n}} \tag{6}$$

The set of all permutations obtained by composing α with itself forms a cyclic group of order n,

$$\underline{\alpha} = \{\alpha^o, \alpha, \alpha^2, \ldots, \alpha^{n-1}\} \tag{7}$$

II. CONFIGURATION GENERATION

This section presents a computer-oriented algorithm to generate all system configurations. Sets of masking generators reduce configuration generation to the union, shift, and masking of binary codes.

P(n) partitions C(S)

We begin by giving a physical interpretation to the elements of *P*(n). For example, $(2,1,1,1) \in P(5)$ means that two terminals are connected by a short-circuit while the other three are free. Similarly, $(3,2) \in P(5)$ means that groupings of three and two terminals are shorted. Notice however that particular terminals are not specified. On the other hand, it is clear that *P*(n) partitions the set of configurations C(S) into "p-classes", $p \in P(n)$, such that configurations in the same p-class have similar terminal groupings. Let C(p) denote a p-class of C(S).

Fact For each $p \in P(n)$, the number of configurations in C(p) is

$$^{\S}C(p) = \frac{n!}{[(f_1!)(f_2!)\ldots(f_k!)]\cdot[(r_1!)\ldots(r_k!)]} \tag{8}$$

where $p = (f_1, f_2, \ldots, f_k)$ and r_i is the number of times a distict f_i appears in p.

For example, $^{\S}C(2,2,1) = \frac{5!}{(2!\ 2!\ 1!)\cdot(2!\ 1!)} = 15$, and the chart below shows how the 52 configurations of a 5-terminal system are partitioned into p-classes:

p	(1,1,1,1,1)	(2,1,1,1)	(2,2,1)	(3,1,1)	(3,2)	(4,1)	(5)
$^{\S}C(p)$	1	10	15	10	10	5	1

Notation Capital letters denote terminals which are shorted together, and small letters denote free terminals. For an n-terminal system, the terminals are labeled 1 through n in a clockwise fashion, and the corresponding configuration is read from left to right, with terminal 1 in the left-most position. Thus the configuration AbAcd means that terminals 1 and 3 are shorted together while terminals 2,4 and 5 are free.

Theorem 1 Let S be an n-terminal system. Each p-class C(p) is the disjoint union of cyclic groups of finite order.

Proof: Each p-class has a finite number of configurations because the total number of configurations is finite.

Assertion For each $p \varepsilon P(n)$, the cyclic group $\underline{\alpha}$ (see(7)) of order n acts on C(p), $\underline{\alpha} \times C(p) \to C(p)$.

To motivate the proof of the assertion, we consider all the configurations in:

$$C(2,1,1) = \{AAbcd, \ dAAbc, \ cdAAb, \ bcdAA, \ AbcdA; \ AbAcd, \ dAbAc, \ cdAbA, \ AcdAb, \ bAcdA\} \tag{9}$$

The elements of C(2,1,1,1) are grouped into two, five-element sets, and within each set the elements are related through the shift operator $\alpha: \underline{5} \to \underline{5}$ as in (7). This argument holds for any $n > 1$ and any $p \varepsilon P(n)$.

The assertion shows that each p-class C(p) is closed under shift operations. We now introduce the relation

$$\simeq \subseteq C(p) \times C(p) \quad \text{such that}$$

$$x \simeq y \quad \text{if and only if } \exists\, i \varepsilon \underline{n} \ni x = \alpha^{i}(y), \qquad x,y \varepsilon C(p)$$

Clearly, $\simeq$ is an equivalence relation which induces a partition of C(p) into disjoint sets. Since the configurations in each $\simeq$-class are equivalent, we can generate each class by specifying a "generating" configuration and the number of times it is to be shifted, ie., its order. Thus, C(2,1,1,1) is the disjoint union of two cyclic groups of order 5.

$$C(2,1,1,1) = \{AAbcd; 5\} \dot{\cup} \{AbAcd; 5\} \tag{10}$$

This completes the proof.

Remark There are special cases of Theorem 1 which are of interest. For p=(1,1,....,1) and p=(n), the p-classes have only one element. For n odd, the order of each $\simeq$-class is either 1 or n, but for n even, the order may be 1, n, or m, where m divides n.

Geometrically speaking, the "degeneracy" m is due to symmetries in polygons with an even number of sides. For example, an $\simeq$-class of (3,1,1,1) is {AbAcAd;2} which corresponds to inscribing an equilateral triangle in a hexagon. Mathematically speaking, the m's are specified by Lagrange's Theorem [6].

Table 1 lists the partitions of n, the configuration generators and their orders for n = 2,3,4,5.

Binary Coding of Configurations

We now proceed to express configuration generators in a binary code. This port code is simpler than the terminal alphabetic code and can easily be handled by the computer.

Since an n-terminal system has n(n-1)/2 port variables, the code has that length. For example, a 5-terminal system has 10 ports which are labeled below in terms of their terminals:

1' =(1,2) 2' =(2,3) 3' =(3,4) 4' =(4,5) 5' =(5,1)
6' =(1,3) 7' =(2,4) 8' =(3,5) 9' =(4,1) 10' =(5,2)

Configuration AAbAc implies that terminals 1,2,4 are shorted. Thus ports 1',7',9' have value 1 while the others have value 0. The code is 10000 01010. Table 1 lists port codes for generators of 2,3,4, and 5-terminal systems.

Notice that ports 1' through 5' above have "distance" one between terminals, while ports 6' through 10' have distance two. A 6-terminal system has 6 ports of distance 1, 6 of distance 2, and 3 of distance 3, making a total of 15 port variables. In general we have the

Fact Let S be an n-terminal system. The distance between terminals (the port distance) is given by the function

$$d: \underline{n} \times \underline{n} \rightarrow Z_n \text{ such that for any } i,j \in \underline{n},$$

$$d(i,j) = \min\{(i-j)\bmod n,\ (j-i)\bmod n\} \qquad (11)$$

where Z_n denotes the integers modulo n.

By grouping port variables in terms of distance we obtain the following important result.

Fact Let S be an n-terminal system which has its port variables grouped in terms of distance. For any configuration $c \in C(S)$ and its corresponding port code b, the permutation $\alpha: \underline{n} \rightarrow \underline{n}$ acting on c induces a similar permutation on each of the distance groupings of b. For example, α^2(AAbAc) = AcAAb implies that 10000 01010 $\mapsto$ 00100 10010.

Masking generators and the generating algorithm.

To completely automate configuration generation we introduce the notions of a weighted partition of n and masking generators. These masks reduce the generating procedure to the union, shift, and masking of binary codes.

Consider a partition $p=(f_1,f_2,\ldots,f_k) \in P(n)$. The number of 1's in the binary codes for the configurations in this p-class is given by

$$\S1_p = \sum_{i=1}^{k} f_i(f_i-1)/2 \qquad (12)$$

In particular, $\S1_{(3,1,\ldots\ldots,1)}=3$. Three ones in the code correspond to inscribing a triangle in an n-sided polygon. Moreover, any two sides automatically specify the third. This idea allows us to use these generators as masks which can be shifted and used to check the validity of a configuration code.

Definition 8 Let S be an n-terminal system. The port variables are grouped in terms of port distances $d_1, d_2, \ldots, d_k$, where d_i denotes distance i.

A *weighted partition of n* is a sequence of k natural numbers $\{w_1,\ldots,w_k\}$ such that

$$\sum_{i=1}^{k} w_i d_i = n \quad \text{and} \quad \sum_{i=1}^{k} w_i = 3$$

The weighted partition set, W(n), is the set of all weighted partitions of n.

Table 2 shows W(n) for n=3,4,5,6,7. The corresponding masking generators and their order are also presented. The exception to the weighted partition of n is the sequence {2,1,0} for n=5,6,7, and the sequence {1,1,1}for n=7.

The computer-oriented configuration generation algorithm.

1. For an n-terminal system S, the number of port variables is $n(n-1)/2$. Group these port variables in terms of distances $d_1,\ldots,d_k$.

2. The partition $p=(1,1,\ldots,1)$ defines the "zero" configuration with all port variables set at zero. The partition $p=(2,1,\ldots,1)$ defines k "basis" configurations of the form.

$$b_i = 00000\ 00000\ 00000\ 10000\ 00000$$

with a 1 in the first position of the d_i^{th} grouping and zeroes elsewhere.

3. Shift the basis configurations by their appropriate orders to obtain all the configurations of $p=(2,1,\ldots,1)$

4. For those partitions between $(2,1,\ldots 1)$ and $(3,1,\ldots,1)$, perform the shifting and union of the basis configurations, taking into account the number of ones in each code (see(12)). The configurations, when masked with the generators of $(3,1,\ldots,1)$ should yield zero correlation.

5. The masks of $(3,1,\ldots,1)$ can be shifted to yield the configurations of that partition.

6. Configuration generators for $p \varepsilon P(n)$ with $(3,1,\ldots,1) \leq p$ can be constructed by using (a) the basis configurations, (b) the number of ones in the code for p, and (c) the masks of $(3,1,\ldots,1)$.

III. THE SYSTEM INTERCONNECTION ALGORITHM

In this section we present an algorithm which allows large systems to be decomposed into smaller subsystems which are analyzed and then recomposed. The overall system's characteristic polynomials are expressed in terms of the subsystems' polynomials. By the same token, the algorithm permits large systems to be constructed from smaller component systems. The advantage of the algorithm is its simplicity and the fact that it can readily be implemented on the computer.

Definition 9 An n-terminal system X is composed of m and p-terminal systems A and B, respectively, written

$$X = A(\underline{1}\ \underline{2}\ \underline{\ \ }\ \underline{\ \ })B(\underline{2}\ \underline{\ \ }\ \underline{4}\ \underline{\ \ }) \tag{13}$$

such that the m and p spaces in the parentheses correspond to the terminals of A and B, while the integers above the spaces indicate how the terminals of A and B are connected to those of X. Integers greater than n denote internal connections. See Table 3 for examples.

Definition 10 Let X be composed of A and B. Given a configuration on X, a set of *external constraints* (EC) and a set of *internal constraints* (IC) are configurations induced on A and B.

Algorithm for computing external and internal constraints.

1. Given a configuration on X, determine which terminals (or groups of terminals) are shorted together.

2. In (13) replace each group of shorted terminals by a common number. Write this number in the appropriate spaces of A and B. Call this equation (13.2).

3. In (13.2) replace distinct numbers by distinct literals. Call this equation (13.3).

4. The external constraints (EC) on A and B are obtained by considering A and B separately in (13.3), and writing the configurations specified by the literals. Notice that repeated literals mean that the corresponding terminals are shorted.

5. The internal constraints (IC) on A and B are obtained by examining repeated literals in A and B and writing the terminal positions in which they appear. Terminal numbers of A and B are separated by a comma. Perform the union of these numbers for A and then for B to obtain all induced terminal shorts. Convert these numbers to port codes.

The system interconnection algorithm.

1. Let X= AB. For a given configuration, c, for X, calculate the internal and external constraints for A and B.

2. The characteristic polynomial for X(c), denoted as $P[X(c)]$ is given by the formula.

$$P[X(c)] = P[A(IC)]\cdot P[B(EC)] + P[A(EC)]\cdot P[B(IC)] \tag{14}$$

where + and . denote addition and multiplication, respectively.

3. If the internal and external constraints are equal for either subsystem, then the system is completely constrained and

$$P[X(c)] = P[A(EC)]\cdot P[B(EC)] \tag{15}$$

4. In some cases a particular configuration may reduce a system's graph to a single node. In this case the characteristic polynomial is set to unity, $P[X(c)] = 1$.

Examples for the system interconnection algorithm are provided in Table 3. Therein, the polynomial notation of (14) and (15) has been dropped; the configurations imply their corresponding polynomials.

IV. ILLUSTRATIVE EXAMPLE

Consider the 3-terminal system S whose graph is shown in Figure 1. The graph has been decomposed into three 3-terminal subsystems A,B, and C. The branch variables denote admittances.

We wish to calculate S(000) by means of the interconnection algorithm. Using the algorithm twice, first on A and B to obtain X = AB, and then on X and C, we obtain S = XC = ABC. The systems have the form of the first entry in Table 3; the formula is S(000) = AB(000)C(001) + AB(010)B(000). The relevant polynomials are:

$A(000) = Y_0Y_1 + Y_0Y_2 + Y_1Y_2$ $\quad B(000) = Y_3Y_4$ $\quad C(000) = Y_5Y_6$

$A(010) = Y_0 + Y_1$ $\quad B(010) = Y_3$ $\quad C(001) = Y_5 + Y_6$

$B(001) = Y_3 + Y_4$

$B(111) = 1$

$$AB(000) = A(000)B(001) + A(010)B(000) = (Y_0Y_1 + Y_0Y_2 + Y_1Y_2)(Y_3 + Y_4) + (Y_0 + Y_1)(Y_3Y_4)$$

$$AB(010) = A(000)B(111) + A(010)B(010) = (Y_0Y_1 + Y_0Y_2 + Y_1Y_2)(1) + (Y_0 + Y_1)(Y_3)$$

$$S(000) = (Y_0Y_1+Y_0Y_2+Y_1Y_2)(Y_3+Y_4)(Y_5+Y_6)+(Y_0+Y_1)(Y_3Y_4)(Y_5+Y_6)$$

$$+(Y_0+Y_1)(Y_3)(Y_5Y_6)+(Y_0Y_1+Y_0Y_2+Y_1Y_2)(1)(Y_5Y_6) \qquad (16)$$

The expansion of (16) yields the twenty-one tree product terms for S(000). The admittance matrix for S with terminal 3 as reference is

$$Y = \frac{1}{S(000)} \begin{bmatrix} S(001) & Y^* \\ Y^* & S(010) \end{bmatrix} \qquad (17)$$

where Y* is the sum of those tree product terms common to both S(001) and S(010). Transfer-function matrices can be obtained from the polynomials and topological methods. The eigenvalues and eigenfunctions of the system can also be obtained from the polynomials.

CONCLUSIONS

The computer-oriented algorithms presented herein completely characterize the configurations of an n-terminal system,and allow the polynomials of a large system to be calculated in terms of the component systems' polynomials.

Further research areas are (1) the computer implementation of the algorithms (2) a formal proof for the interconnection algorithm, (3) complexity bounds for system graphs (In certain cases, a configuration induces a graph which is symmetrical, and the algorithm looses some tree product terms), and (4) the application of the algorithm to problems in Graph Theory, Switching Theory, Mechanics, and Control Theory.

REFERENCES

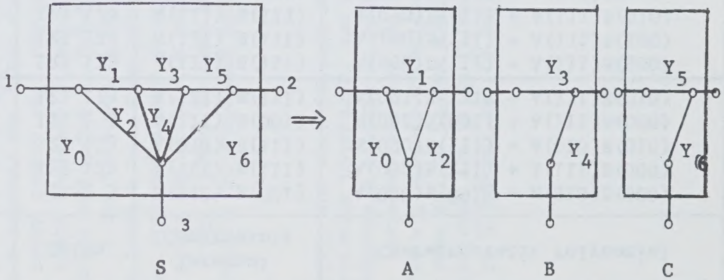
1. R.E.D. Bishop and D.C. Johnson, The Mechanics of Vibration, Cambridge, at the University Press, 1960.
2. M. Javid and E. Brenner, Analysis, Transmission, and Filtering of Signals, McGraw Hill, New york, 1963.
3. C.T. Chen and C.A. Desoer, "Controllability of Composite Systems", IEEE Trans. Auto. Control, Vol. AC-12, pp. 402-409, August, 1967.
4. A.S. Krajewski and W.W. Happ, "Deviations- A New Calculus for Electronic Circuits", ROZPRAWY ELEKTROTECHNICZNE, Vol.XII, No. 4, pp. 675-708, 1966.
5. S.P. Chan, Introductory Topological Analysis of Electrical Networks, Holt, Rinehart, and Winston, Inc., New York, 1969.
6. S. MacLane and G. Birkhoff, Algebra, The Macmillan Co., New York, 1967.

n	Partition	Configuration Generator	Order	Binary Code
2	(1,1)	ab	1	0
	(2)	AA	1	1
3	(1,1,1)	abc	1	000
	(2,1)	AAb	3	100
	(3)	AAA	1	111
4	(1,1,1,1)	abcd	1	0000 00
	(2,1,1)	AAbc	4	1000 00
		AbAc	2	0000 10
	(2,2)	AABB	2	1010 00
		ABAB	1	0000 11
	(3,1)	AAAb	4	1100 10
	(4)	AAAA	1	1111 11
5	(1,1,1,1,1)	abcde	1	00000 00000
	(2,1,1,1)	AAbcd	5	10000 00000
		AbAcd	5	00000 10000
		AABBc	5	10100 00000
	(2,2,1)	ABABc	5	00000 11000
		ABBAc	5	01000 00010
	(3,1,1)	AAAbc	5	11000 10000
		AAbAc	5	10000 01010
	(3,2)	AAABB	5	11010 10000
		AABAB	5	10000 01110
	(4,1)	AAAAb	5	11100 11010
	(5)	AAAAA	1	11111 11111

Table 1: Configuration Generator Codes

n	$\S_C(3,1,..,1)$	W(n) d_1	d_2	d_3	Masking Generators	Order
3	1	3	0	0	111	1
4	4	2	1	0	1100 10	4
5	10	2	1	0	11000 10000	5
		1	2	0	10000 01010	5
6	20	2	1	0	110000 100000 000	6
		1	1	1	100000 010000 100	6
		1	1	1	100000 000010 100	6
		0	3	0	000000 101010 000	2
7	35	2	1	0	1100000 1000000 0000000	7
		1	1	1	1000000 0100000 1000000	7
		1	0	2	1000000 0000000 0100100	7
		0	2	1	0000000 1100000 0000100	7
		0	2	1	0000000 1000010 0010000	7

Table 2: W(n) and Masking Generators



A 3-terminal system S,subsystems A, B, and C
Figure 1

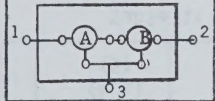
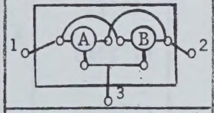
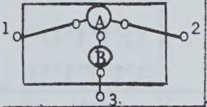
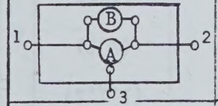
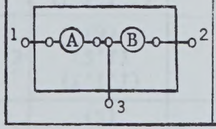
Composed System	Test	Literals	External Constraints	Union	Internal Constraints	Characteristic Polynomial
X = A(143)B(423)  	X(000)	A(abc)B(bdc)	A(000) B(000)	23,1 3	A(010) B(001)	A(000)B(001) + A(010)B(000)
	X(100)	A(abc)B(bac)	A(000) B(000)	123,123	A(111) B(111)	A(000)B(111) + A(111)B(000)
	X(010)	A(abc)B(bcc)	A(000) B(010)	23,123	A(010) B(111)	A(000)B(111) + A(010)B(010)
	X(001)	A(aba)B(bca)	A(001) B(000)	123,1 3	A(111) B(001)	A(001)B(001) + A(111)B(000)
	X(111)	A(aba)B(baa)	A(001) B(010)	123,123	A(111) B(111)	A(001)B(111) + A(111)B(010)
X = A(123)B(123)  	X(000)	A(abc)B(abc)	A(000) B(000)	123,123	A(111) B(111)	A(000)B(111) + A(111)B(000)
	X(100)	A(aab)B(aab)	A(100) B(100)	123,123	A(111) B(111)	A(100)B(111) + A(111)B(100)
	X(010)	A(abb)B(abb)	A(010) B(010)	123,123	A(111) B(111)	A(010)B(111) + A(111)B(010)
	X(001)	A(aba)B(aba)	A(001) B(001)	123,123	A(111) B(111)	A(001)B(111) + A(111)B(001)
	X(111)	A(aaa)B(aaa)	A(111) B(111)	123,123	A(111) B(111)	A(111)B(111)
X = A(124)B(43)  	X(000)	A(abc)B(cd)	A(000) B(0)	3,1	A(000) B(0)	A(000)B(0)
	X(100)	A(aac)B(cd)	A(100) B(0)	12 ,	A(100) B(0)	A(100)B(0)
	X(010)	A(abc)B(cb)	A(000) B(0)	23,12	A(010) B(1)	A(000)B(1) + A(010)B(0)
	X(001)	A(abc)B(ca)	A(000) B(0)	1 3,12	A(001) B(1)	A(000)B(1) + A(001)B(0)
	X(111)	A(aac)B(ca)	A(100) B(0)	123,12	A(111) B(1)	A(100)B(1) + A(111)B(0)
X = A(123)B(12)  	X(000)	A(abc)B(ab)	A(000) B(0)	12 ,12	A(100) B(1)	A(000)B(1) + A(100)B(0)
	X(100)	A(aac)B(aa)	A(100) B(1)	12 ,12	A(100) B(1)	A(100)B(1)
	X(010)	A(abb)B(ab)	A(010) B(0)	123,12	A(111) B(1)	A(010)B(1) + A(111)B(0)
	X(001)	A(aba)B(ab)	A(001) B(0)	123,12	A(111) B(1)	A(001)B(1) + A(111)B(0)
	X(111)	A(aaa)B(aa)	A(111) B(1)	123,12	A(111) B(1)	A(111)B(1)
X = A(13)B(23)  	X(000)	A(ab)B(cb)	A(0) B(0)	2 , 2	A(0) B(0)	A(0)B(0)
	X(100)	A(ab)B(ab)	A(0) B(0)	12 ,12	A(1) B(1)	A(0)B(1) + A(1)B(0)
	X(010)	A(ab)B(bb)	A(0) B(1)	,1 3	A(0) B(1)	A(0)B(1)
	X(001)	A(aa)B(ba)	A(1) B(0)	12 ,	A(1) B(0)	A(1)B(0)
	X(111)	A(aa)B(aa)	A(1) B(1)	12 ,12	A(1) B(1)	A(1)B(1)

Table 3: The System Interconnection Algorithm; Examples.

INFLATION IN LINEAR DYNAMICAL SYSTEMS

MICHAEL K. SAIN
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana 46556

ABSTRACT

Let A be a square matrix having a proper value p. An inflation of A is a larger square matrix, the non-p proper values of which are precisely those of A; and a deflation of A is a smaller square matrix of the same restriction. When A defines the incremental behavior in systems $\dot{x} = Ax$, or $x_{k+1} = Ax_k$, parallel notions follow naturally for the systems also. Inflationary and delationary techniques can be used to exhibit a decomposition in which the p-values of a dynamical system are aggregated separately from the non-p ones. The decomposition works for systems over any field.

INTRODUCTION

In finding the zeros of a polynomial, a common procedure is to remove the factor corresponding to a zero as soon as the zero is found. This process assumes, of course, that the polynomial has a zero in the field E of its coefficients. In case it does not, then it is always possible to find an extension field of E in which the polynomial does have a zero; in fact [1, p. 30], for a given polynomial and a given E, there exists a minimal extension field E′, called a splitting field, in which the polynomial has a number of zeros, with multiplicities properly counted, equal to its degree, and which is unique up to an isomorphism. The sequel makes use of E′ as a convenience in making certain arguments only; the results are readily applicable in the original field E, provided that there is at least one zero in E.

A polynomial of considerable interest in linear dynamical system theory is the characteristic polynomial $|vI - A|$, where I is a diagonal matrix, compatible with A, of identity elements over E, and $|\cdot|$ denotes the determinant of an enclosed square matrix. The zeros of such polynomials, of course, play a central role in the dynamic response of dynamical systems of types

$$\dot{x} = Ax \ , \ t \geq 0 \ , \tag{1}$$

or

$$x_{k+1} = Ax_k \ , \ k = 0, 1, 2, 3, \ldots \ , \tag{2}$$

where x(t), or x_k, respectively, takes values in E^n, the n-fold Cartesian product of E. In this context, zeros of $|vI - A|$ will be called proper values of A. It is quite well known, to be sure, that p is a zero of $|vI - A|$ if and only if there is a nonzero e in E^n such that

$$Ae = pe \ . \tag{3}$$

The idea of dividing out known factors of a polynomial extends naturally to removal of known proper values of a matrix and is not new in the mathematical literature. Householder [2, p. 168] terms such a process a

---

This research is supported by the National Science Foundation under the Grant GK-13618.

deflation. Application of the idea to decomposition of dynamical systems seems relatively novel at this writing, however; and such applications seem to be especially interesting because of the presence of dependent variables in both members of (1) and (2). This is not the case, obviously, in solving the linear algebraic equation $Ax = b$. The significance of these points should become clearer as the ensuing decomposition is pursued.

NOTATION

For purposes of generality in the sequel, the representations (1) and (2) are now extended to include inputs $u(t)$, or u_k, respectively, and corresponding outputs $y(t)$, or y_k. Inputs are understood to take values in E^m, outputs in E^r. The extended equations are then

$$\dot{x} = Ax + Bu \ , \ t \geq 0 \ , \tag{4}$$

$$y = Cx \ , \tag{5}$$

or

$$x_{k+1} = Ax_k + Bu_k \ , \ k = 0, 1, 2, 3, \ldots \ , \tag{6}$$

$$y_k = Cx_k \ . \tag{7}$$

It will be convenient to develop the decomposition simultaneously for both (4)-(5) and (6)-(7), with the understanding that (4)-(5) is not a workable description on a field E for which the derivative is not well defined. As a common notation for the two cases, the remainder of the paper uses the sixtuple

$$[C, A, B; y, x, u] \ . \tag{8}$$

Use of the notation (8) should be understood as a shorthand for the equation pair (4)-(5) or (6)-(7), as the case may be. Note that (8) is particularly effective in indicating changes by linear transformation; for example, let $z = Tx$, where T is nonsingular, and obtain

$$[CT^{-1}, TAT^{-1}, TB; y, z, u] \ . \tag{9}$$

Of course, corresponding transformations must also be carried out upon the initial conditions of (4) or (6); these steps will not be made explicit in the notation, but it should be understood that they are implied.

ONE-STEP DECOMPOSITION: P = ZERO

Consider the system $[C, A_1, B; y, x, u]$ with A_1 having a zero proper value. The matrix A_1 has the factorization F_1G_1, where F_1 and G_1 both have the same rank as A_1, and where F_1 has a number of columns equal to the rank of A_1. Let H_1 be any matrix such that

$$\begin{bmatrix} G_1 \\ H_1 \end{bmatrix} = T_1 \tag{10}$$

has an inverse; such a matrix, of course, always exists. Then the transformation $q_1 = T_1x$ accomplishes the first step in the desired decomposition, which is to say that the system $[CT_1^{-1}, T_1A_1T_1^{-1}, T_1B; y, q_1, u]$ aggregates a block of zero proper values separately from the remaining proper values. The conclusion follows directly from the identity

$$\begin{bmatrix} G_1 \\ H_1 \end{bmatrix} F_1 G_1 \begin{bmatrix} G_1 \\ H_1 \end{bmatrix}^{-1} = T_1 A_1 T_1^{-1} = \begin{bmatrix} G_1F_1 & 0 \\ H_1F_1 & 0 \end{bmatrix} , \tag{11}$$

which itself is an easy consequence of the simpler identity

$$G_1 \begin{bmatrix} G_1 \\ H_1 \end{bmatrix}^{-1} = [I \ \ 0] . \tag{12}$$

If G_1F_1 has an inverse, then all the zero proper values have been aggregated separately from the nonzero ones in the first step of decomposition and the process terminates. If G_1F_1 does not have an inverse, then only a portion of the zero proper values have been separated from the rest, and another step of decomposition must be undertaken.

TWO-STEP DECOMPOSITION: P = ZERO

Define the matrix G_1F_1 to be A_2; since A_2 does not have an inverse, it has at least one zero proper value. Accordingly, A_2 has the factorization F_2G_2, where F_2 and G_2 both have the same rank as A_2, and where F_2 has a number of columns equal to the rank of A_2. Let H_2 be any matrix such that

$$\begin{bmatrix} G_2 \\ H_2 \end{bmatrix} \tag{13}$$

has an inverse; as in the first step above, such a matrix always exists. Then

$$\begin{bmatrix} \begin{pmatrix} G_2 \\ H_2 \end{pmatrix} & 0 \\ 0 & I \end{bmatrix} = T_2 \tag{14}$$

also has an inverse, and the transformation $q_2 = T_2q_1$ accomplishes the second step in the decomposition, giving a system described by the six-tuple $[CT_1^{-1}T_2^{-1},\ T_2T_1A_1T_1^{-1}T_2^{-1},\ T_2T_1B;\ y,\ q_2,\ u]$. The identity matrix in the transformation of equation (14) is chosen so that T_2 is compatible with q_1, which in turn assures that q_2 takes values in E^n, as did q_1 also. In order to exhibit the additional block of zero proper values which has been aggregated by the second step of decomposition, observe that

$$T_2T_1A_1T_1^{-1}T_2^{-1} = T_2 \begin{bmatrix} G_1F_1 & 0 \\ H_1F_1 & 0 \end{bmatrix} T_2^{-1} , \tag{15}$$

by virtue of (11); from (14) and the identity

$$G_2 \begin{bmatrix} G_2 \\ H_2 \end{bmatrix}^{-1} = [I \ \ 0] , \tag{16}$$

the right member of (15) becomes

$$\left[\begin{array}{ccc} G_2F_2 & 0 & 0 \\ H_2F_2 & 0 & 0 \\ \multicolumn{2}{l}{\left(H_1F_1 \begin{pmatrix} G_2 \\ H_2 \end{pmatrix}^{-1} \right)} & 0 \end{array}\right] \tag{17}$$

If G_2F_2 has an inverse, then all the zero proper values have been aggregated separately from the nonzero ones by the first two steps of decomposition, and the process terminates. If G_2F_2 does not have an inverse, then only a second portion of the zero proper values have been separated from the rest, and further steps of decomposition must be undertaken.

In the latter event, the matrix in parentheses in (17) is a bit inconvenient for further steps; one workable way to overcome the difficulty is to partition it conformably with the columns above it. This can be accomplished in a straightforward manner by so partitioning the enclosed matrix inverse according to the scheme

$$\begin{bmatrix} G_2 \\ H_2 \end{bmatrix}^{-1} = [Q_1 \quad R_1] \ . \tag{18}$$

The array (17) can then be rewritten as

$$\begin{bmatrix} G_2F_2 & 0 & 0 \\ H_2F_2 & 0 & 0 \\ H_1F_1Q_1 & H_1F_1R_1 & 0 \end{bmatrix} \ . \tag{19}$$

I-STEP DECOMPOSITION: P = ZERO

Suppose that, after i-1 steps of decomposition, the matrix $G_{i-1}F_{i-1}$ does not have an inverse; define it to be A_i, and note that A_i must have at least one zero proper value. Accordingly, A_i has the factorization F_iG_i, where F_i and G_i both have the same rank as A_i, and where F_i has a number of columns equal to the rank of A_i. Let H_i be any matrix such that

$$\begin{bmatrix} G_i \\ H_i \end{bmatrix} \tag{20}$$

has an inverse; such a matrix always exists. Then

$$\begin{bmatrix} G_i & 0 \\ H_i & \\ & \\ 0 & I \end{bmatrix} = T_i \tag{21}$$

also has an inverse, and the transformation $q_i = T_iq_{i-1}$ accomplishes the ith step in the decomposition, giving a system described by the sixtuple

$$[CT_1^{-1}T_2^{-1}\ldots T_i^{-1},\ T_i\ldots T_2T_1A_1T_1^{-1}T_2^{-1}\ldots T_i^{-1},\ T_i\ldots T_2T_1B;\ y,\ q_i,\ u] \ . \tag{22}$$

Analogously to (18), in which Q_1 has the same number of columns as F_2, the inverse of (20) is partitioned in the manner $[Q_{i-1} \quad R_{i-1}]$, so that Q_{i-1} has the same number of columns as F_i.

For i not less than two, it is a quite straightforward recursive construction to show that the general forms of the three matrices in (22) can be given as follows:

$$CT_1^{-1}T_2^{-1}\ldots T_i^{-1} = C[Q_0Q_1Q_2\ldots Q_{i-1} \quad Q_0Q_1\ldots Q_{i-2}R_{i-1} \quad Q_0Q_1\ldots Q_{i-3}R_{i-2} \ldots$$
$$\ldots Q_0Q_1R_2 \quad Q_0R_1 \quad R_0] \ , \tag{23}$$

where the inverse of (10) is written $[Q_0 \quad R_0]$, with Q_0 having the same number of columns as F_1;

$$T_i\ldots T_2T_1A_1T_1^{-1}T_2^{-1}\ldots T_i^{-1} = \begin{bmatrix} G_iF_i & 0 & 0 \\ H_iF_i & 0 & 0 \\ V_i & W_i & 0 \end{bmatrix} , \tag{24}$$

for

$$V_i = \begin{bmatrix} H_{i-1}F_{i-1}Q_{i-1} \\ H_{i-2}F_{i-2}Q_{i-2}Q_{i-1} \\ H_{i-3}F_{i-3}Q_{i-3}Q_{i-2}Q_{i-1} \\ \cdots \quad \cdots \\ H_2F_2Q_2Q_3\cdots Q_{i-1} \\ H_1F_1Q_1Q_2\cdots Q_{i-1} \end{bmatrix} \tag{25}$$

and

$$W_i = X_iY_iZ_i \ , \tag{26}$$

where

$$X_i = \begin{bmatrix} H_{i-1}F_{i-1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & H_{i-2}F_{i-2} & 0 & \cdots & 0 & 0 \\ 0 & 0 & H_{i-3}F_{i-3} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & H_2F_2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & H_1F_1 \end{bmatrix} , \tag{27}$$

$$Y_i = \begin{bmatrix} I & 0 & 0 & \cdots & 0 & 0 \\ Q_{i-2} & I & 0 & \cdots & 0 & 0 \\ Q_{i-3}Q_{i-2} & Q_{i-3} & I & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ Q_2Q_3\cdots Q_{i-2} & Q_2\cdots Q_{i-3} & Q_2\cdots Q_{i-4} & \cdots & I & 0 \\ Q_1Q_2\cdots Q_{i-2} & Q_1\cdots Q_{i-3} & Q_1\cdots Q_{i-4} & \cdots & Q_1 & I \end{bmatrix} , \tag{28}$$

$$Z_i = \begin{bmatrix} R_{i-1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & R_{i-2} & 0 & \cdots & 0 & 0 \\ 0 & 0 & R_{i-3} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & R_2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & R_1 \end{bmatrix} ; \tag{29}$$

$$T_i\cdots T_2T_1B = \begin{bmatrix} G_iG_{i-1} \cdots G_1 \\ H_iG_{i-1}G_{i-2} \cdots G_1 \\ H_{i-1}G_{i-2}G_{i-3} \cdots G_1 \\ \cdots \quad \cdots \\ H_2G_1 \\ H_1 \end{bmatrix} B \tag{30}$$

To complete the interpretation of (28), define $Y_2 = I$.

DECOMPOSITION AS A DEFLATIONARY PROCESS

The process of decomposition exhibited in the previous sections is an example of a deflationary process; moreover, the inverse process of reconstructing the original system from the decomposition is an example of an inflationary process. These terms can be made more precise, as follows.

Definition 1

Let A be a square matrix having a proper value p. An inflation of A is a larger square matrix, the non-p proper values of which are precisely those of A; and a deflation of A is a smaller square matrix of the same restriction.

Under Definition 1, the process of decomposition foregoing is deflationary, which is the content of the following theorem.

Theorem 1

The matrix A_{i+1}, appearing as G_iF_i in (24) for an i-step decomposition in the sense of the previous section, is a deflation of the matrix A_1 relative to the proper value zero.

Proof of Theorem 1:

To prove the stated result, it is only necessary to establish that the nonzero proper values of the matrix G_iF_i are precisely the same as the nonzero proper values of the matrix F_iG_i, for this is how the construction of A_{i+1} is defined: A_1 is factored into the product F_1G_1, G_1F_1 is factored into the product F_2G_2, G_2F_2 is factored into the product F_3G_3, and so forth.

The elements of A_1 take their values in the field E under the assumptions of this paper; in order that the nonzero proper values of G_iF_i be completely defined, and so that the total number of proper values be n in number, the field E is minimally extended to the splitting field E′. Let s be a nonzero element in E′, and let e be a nonzero element of $(E')^n$; suppose that s is a nonzero proper value of G_iF_i and that e is its corresponding proper vector. By definition, then,

$$G_iF_ie = se \quad . \tag{31}$$

Premultiplication of both members of (31) by F_i gives

$$F_iG_i(F_ie) = s(F_ie) \quad ; \tag{32}$$

now F_ie cannot vanish, for otherwise s would have to be zero in (31), and therefore s is a nonzero proper value of F_iG_i also, with corresponding proper vector F_ie. An analogous argument establishes that any nonzero proper value of F_iG_i is also a nonzero proper value of G_iF_i. QED

Remark

It may be of interest to point out that the argument used in proving Theorem 1 is the fundamental principle underlying one of the most frequently used matrix inverse identities in control theory [3, p. 109].

In the sense of Definition 1, an examination of (11), (19), (24), and (26)-(29) establishes also the following result, which is inverse to Theorem 1 in an appropriate sense.

Theorem 2

The matrix $T_i T_{i-1} \ldots T_2 T_1 A_1 T_1^{-1} T_2^{-1} \ldots T_{i-1}^{-1} T_i^{-1}$ is an inflation of the matrix A_{i+1}, provided that the ith step of decomposition is not the terminal step, and relative to the proper value zero.

Remark

The proviso in Theorem 2 assures that A_{i+1} has at least one zero proper value, so that Definition 1 applies.

Consider now equations (4) and (6), which can be denoted generically by

$$[A,\ B;\ x,\ u] \tag{33}$$

in a manner similar to (8). By using the definitions for inflation and deflation of A, it is possible to bring forth the same concepts for associated systems.

Definition 2

Let the matrix A_a be an inflation of the matrix A_b; then for appropriately dimensioned matrices B_a and B_b the system $[A_a, B_a; x_a, u]$ is an inflation of the system $[A_b, B_b; x_b, u]$. If a matrix A_a is a deflation of the matrix A_b, then for appropriately dimensioned matrices B_a and B_b the system $[A_a, B_a; x_a, u]$ is a deflation of the system $[A_b, B_b; x_b, u]$.

From this definition and from Theorems 1 and 2 follow two corollaries.

Corollary 1

The system $[A_{i+1}, G_i G_{i-1} \ldots G_1 B; G_i G_{i-1} \ldots G_1 x, u]$ is a deflation of the system $[A_1, B; x, u]$ relative to the proper value zero.

Proof of Corollary 1:

Although the corollary follows directly from Theorem 1 and Definition 2 it is useful to point out that the second entry in the first fourtuple of the corollary may be read from the first row of (30); the third entry in the same fourtuple follows similarly. In order to make this last issue intuitively more clear, however, consider the following sequence of matrix operations on equation (4), which is

$$\dot{x} = F_1 G_1 x + Bu \quad . \tag{34}$$

Premultiply both members by G_1 to obtain

$$(G_1 \dot{x}) = (G_1 F_1)(G_1 x) + G_1 Bu \quad ; \tag{35}$$

factor so that $G_1 F_1 = F_2 G_2$, and premultiply both members by G_2 to yield

$$(\dot{G_2 G_1 x}) = (G_2 F_2)(G_2 G_1 x) + G_2 G_1 Bu \quad . \tag{36}$$

Continuation up to the ith step provides an alternate way to prove the corollary.

Corollary 2

The system $[T_i \ldots T_2 T_1 A_1 T_1^{-1} T_2^{-1} \ldots T_i^{-1}, T_i \ldots T_2 T_1 B; q_i, u]$ is an inflation of the system $[A_{i+1}, G_i G_{i-1} \ldots G_1 B; G_i G_{i-1} \ldots G_1 x, u]$, provided that the ith step of decomposition is not the terminal step, and relative to the proper value zero.

DECOMPOSITION WHEN P $\neq$ ZERO

In the case when p is nonzero, simply rewrite A_1 as $[(A_1 - pI) + pI]$ so that the new matrix $(A_1 - pI)$ has zero proper values. Then, by making use of the identity

$$\begin{aligned} &T_i \ldots T_2 T_1 [(A_1 - pI) + pI] T_1^{-1} T_2^{-1} \ldots T_i^{-1} = \\ &\quad = T_i \ldots T_2 T_1 (A_1 - pI) T_1^{-1} T_2^{-1} \ldots T_i^{-1} + pI \quad , \end{aligned} \tag{37}$$

the same decomposition may be applied to $(A_1 - pI)$ as was developed in the earlier sections for p = zero. The only change that is necessary in the prior results for this case is to add p to every main diagonal element of the matrices (11), (19), and (24).

BINARY EXAMPLE

As an example of the decompositions, consider an equation of the type (6)-(7) over the binary field. The matrices B and C will not be explicitly stated in numeric terms, but the matrix A is assumed to be

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad . \tag{38}$$

Matrix (38) has two zero proper values and one unity proper value; further, it has rank two. Accordingly, a first factorization $F_1 G_1$ is chosen, with

$$F_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \quad , \quad G_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad ; \tag{39}$$

The product $G_1 F_1 = A_2$ is calculated to be

$$\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \tag{40}$$

which has an $F_2 G_2$ factorization of the form

$$F_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad , \quad G_2 = [1 \;\; 1] \quad . \tag{41}$$

The final product $A_3 = G_2 F_2$ then becomes [1]. Matrices

$$H_1 = [0 \;\; 0 \;\; 1] \quad , \quad H_2 = [0 \;\; 1] \tag{42}$$

can then be chosen, with the result that

$$Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} , \quad R_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} , \quad Q_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} , \quad R_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad . \tag{43}$$

The defining matrices of the decomposed system then are calculated to be

$$CT_1^{-1}T_2^{-1} = C[Q_0Q_1 \quad Q_0R_1 \quad R_0] = C \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} , \tag{44}$$

$$T_2T_1A_1T_1^{-1}T_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} , \tag{45}$$

$$T_2T_1B = \begin{bmatrix} G_2G_1 \\ H_2G_1 \\ H_1 \end{bmatrix} B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} B . \tag{46}$$

It is also possible, of course, to decompose relative to the unity proper value. In this instance,

$$A_1 - I = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} , . \tag{47}$$

which can be factored into F_1G_1 where

$$F_1 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} , \quad G_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} . \tag{48}$$

The reverse product G_1F_1 becomes

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{49}$$

which then terminates the process. H_1 can be chosen $[0 \quad 1 \quad 0]$, so that

$$Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} , \quad R_0 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} . \tag{50}$$

The defining matrices

$$CT_1^{-1} = C[Q_0 \quad R_0] = C \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} , \tag{51}$$

$$T_1B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} , \tag{52}$$

$$T_1(A_1 - I)T_1^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} , \tag{53}$$

are complete except for (53), to which must be added the identity matrix so that

$$T_1A_1T_1^{-1} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad . \tag{54}$$

SUMMARY

The procedure of decomposition which has been exhibited in these sections is oriented toward separating the proper values of a matrix into two groups, one containing all the values which are equal, say, to p, and the other containing all the rest. A technique of deflation provides the decomposition in a step-by-step fashion. Since the case in which p is nonzero can be reduced to a related problem in which it is zero, the steps have been described in detail for that case.

A feature of the deflationary decomposition presented is its ability to treat simultaneously both differential and difference equations, over any field for which the equations are well defined. In emphasis of this point, a detailed binary example has been described.

Prerequisite ideas for the decomposition are those of proper value, rank, and matrix factorization. Because it proceeds in a step-by-step construction, the decomposition proves its own existence; this feature may be compared, for example, with the proof of the Jordan form existence. It should be observed also that the decreases in rank occurring in A_1, A_2,..., are precisely related to the Jordan block structure of the matrix A, and can be used to determine that structure if so desired. The orderly development of the decomposition block structure is another advantage conceptually over techniques which do not array diagonal matrix blocks in any particular order. Essentially, the deflationary decomposition procedure seems to possess some attractive advantages in terms of what can be accomplished relative to the knowledge which has to be assumed.

Some of the interesting applications of the technique must be omitted here for lack of space. Consider, for example, the problem of integrating the matrix exponential, that is, of finding an explicit form for the expression $\int(\exp At)dt$. If A is nonsingular, the answer is well known. If, however, A is singular, then a closed form expression can still be obtained with the techniques brought forth in this paper.

REFERENCES

1. Emil Artin, Galois Theory. Notre Dame, Indiana: University of Notre Dame Press, 1959.
2. A. S. Householder, The Theory of Matrices in Numerical Analysis. New York: Blaisdell Publishing Company, 1964.
3. M. K. Sain, ''On the Control Applications of a Determinant Equality Related to Eigenvalue Computation,'' IEEE Transactions on Automatic Control, Vol. AC-11, pp. 109-111, January 1966.

A SIMPLIFICATION OF THE STATE REGULATOR PROBLEM

W. L. McDANIEL, JR.
Mississippi State University
State College, Mississippi

V. L. DUGAN
Sandia Laboratory
Albuquerque, New Mexico

An appreciable amount of time and effort may be conserved when finding the optimal control law for a linear, time-invariant state-regulator (or output regulator) with a free end time if a simple point is noticed with regard to the solution of the Riccati equation.

Consider a completely controllable, linear, time-invariant, n^{th} order system

$$\dot{\underline{x}} = A\underline{x} + B\underline{u} \quad , \tag{1}$$

and a performance index

$$P.I. = (1/2)\int_0^\infty (\underline{x}^TQ\underline{x} + \underline{u}^TP\underline{u})\, dt \quad , \tag{2}$$

where Q is at least positive semidefinite and P is positive definite. As has been shown in the literature[1,2], the optimal control law for this problem is

$$\underline{u}^o = -P^{-1}B^TR\underline{x} \quad , \tag{3}$$

where R is a constant, positive definite, symmetric matrix which satisfies the Riccati equation

$$A^TR + RA - RBP^{-1}B^TR + Q = 0 \quad . \tag{4}$$

The common practice is to obtain the matrix R by solving the n(n+1)/2 simultaneous, *non-linear*, algebraic equations provided by (4). This solution is then substituted into equation (3) to define the optimal control law. If the control law is written

$$\underline{u}^o(\underline{x}) = -K^T\underline{x} \quad , \tag{5}$$

where

$$K^T = P^{-1}B^TR \quad , \tag{6}$$

the Riccati equation may also be expressed as

$$A^TR + RA - RBK^T + Q = 0 \tag{7}$$

or as

$$R(A-BK^T) + A^TR = -Q \quad . \tag{8}$$

One may now notice that by using the matrix equation (8) the n(n+1)/2 elements of R may be determined in terms of the k_{ij} by solving n(n+1)/2 *linear* equations. Since P^{-1} and B^T are known and since the elements of R may now be expressed in terms of the elements of K, the elements of K are found by solving the nm (m is the order of the control vector) *non-linear* equations[3] resulting from

$$K^T = P^{-1}B^TR \quad . \tag{9}$$

Obviously, there is no savings unless nm < n(n+1)/2, where the two sides of the inequality represent the number of non-linear equations to be solved by the suggested method or by the strict solution of the Riccati equation, respectively.

The real worth of the suggested method is made manifest with the single-input system for which m = 1 and $\underline{u}$ = u. For this case a maximum of only n non-linear equations need be solved using the suggested approach.

Again this compares with the maximum of n(n+1)/2 non-linear equations which may arise in the straight-forward solution.

For further clarification consider a single-input system

$$\dot{\underline{x}} = A\underline{x} + \underline{b}u \tag{10}$$

and a performance index

$$\text{P.I.} = (1/2)\int_0^\infty (\underline{x}^T Q\underline{x} + pu^2)dt \ . \tag{11}$$

Let $\underline{b}$ be restricted to

$$\underline{b} = \text{col}[0 \quad 0 \text{ --- } 0 \quad 1] \ . \tag{12}$$

The equation

$$R(A-\underline{bk}^T) + A^T R = -Q \tag{13}$$

is first solved to find the elements of R as functions of the elements of $\underline{k}$. Secondly, the equation

$$\underline{k}^T = (1/p)\ \underline{b}^T R \tag{14}$$

or equivalently,

$$[k_1 \ k_2 \text{----} \ k_n] = (1/p)[r_{n1} \ r_{n2} \text{ ---- } r_{nn}] \tag{15}$$

may be solved for the k_i's by equating corresponding elements of the two vectors.

In many cases the above method allows the hand solution of a problem which may otherwise require a machine solution. Also it may allow the determination of the k_i's as functions of parameters which have no single numerical specification.

REFERENCES

1. R. E. Kalman, "When is a Linear Control System Optimal?", Transactions of the ASME - Journal of Basic Engineering, pp. 53-55, March, 1964.

2. M. Athans and P. L. Falb, Optimal Control. New York: McGraw-Hill, 1966, ch. 9.

3. D. G. Schultz and J. L. Melsa, State Functions and Linear Control Systems. New York: McGraw-Hill, 1967, Ch. 7.

SENSITIVITY MINIMIZATION IN THE DESIGN OF LINEAR MULTIVARIABLE FEEDBACK SYSTEMS *

M. G. STRINTZIS
University of Pittsburgh
Pittsburgh, Pa. 15213

B. LIU
Princeton University
Princeton, N.J. 08540

ABSTRACT

This paper examines multivariable r-input, n-output linear systems whose unknown, uncertain or varying parameters can be considered random. A linear feedback controller is determined so that an appropriate sensitivity measure is minimized under a stability constraint. The analysis produces an explicit formula for the physically realizable optimum controller.

1. INTRODUCTION

The problem examined in this paper can be stated as follows: Consider a linear dynamic process to be controlled called the "Plant" F and suppose that the characteristics of the process are either varying or not fully known. We seek to determine a linear feedback controller H such that, for the given input $\underline{u}$, the difference between the actual output $\underline{y}$ and the desired output $\underline{y}_o$ is minimized according to some criterion. The problem has been thoroughly examined [1] for the single-input, single-output case. A convenient measure of the system's sensitivity to errors in the estimation of the unknown parameters or to parameter variation is easily defined in this case. For example, if T is the overall system transfer function and $\underline{x}$ is the unknown or varying vector, the "Bode Sensitivity" $\frac{dT(s,\underline{x})}{d\underline{x}} \cdot \frac{\underline{x}}{T(s,\underline{x})}$ provides a measure depending on the system characteristics only and independent of the input. In the multivariable case, however, each output depends on more than one input. Thus the sensitivity examination of multivariable systems is meaningful only when made in relation to one specific input or to an appropriately weighted group of possible inputs.

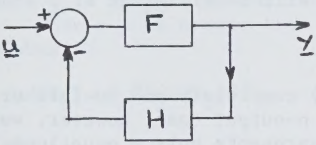


FIG. 1

Another difficulty associated with the multivariable case is that simultaneous reduction of the variations of all outputs y_i is not generally possible. We must then weigh the importance of the variation of each output component and define an index of system performance accordingly. An index that would be appropriate for most problems is

$$I(\underline{x}) = \int_{-\infty}^{+\infty} [\underline{y}(t,\underline{x}) - \underline{y}_o(t)]^* \, \Psi [\underline{y}(t,\underline{x}) - \underline{y}_o(t)] \, dt$$

where $y(t,\underline{x})$ is the system output that corresponds to parameter vector $\underline{x}$, $\underline{y}_o(t)$ is the desired output and Ψ is a positive definite weighing matrix. The asterisk * denotes the complex conjugate of the transpose of the vector $\underline{y}-\underline{y}_o$. By using the Parseval relation, $I(\underline{x})$ can be written as:

$$I(\underline{x}) = \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} [\underline{y}(s,\underline{x}) - \underline{y}_o(s)]_* \Psi [\underline{y}(s,\underline{x}) - \underline{y}_o(s)] \, ds$$

where $\underline{y}(s,\underline{x})$ and $\underline{y}_o(s)$ are the bilateral Laplace transforms of $\underline{y}(t,\underline{x}), \underline{y}_o(t)$ respectively and $[\underline{y}(s,\underline{x}) - \underline{y}_o(s)]*$ is a short notation for $[\underline{y}(-s,\underline{x}) - \underline{y}_o(-s)]^T$.

---

* The research in this paper was performed at Princeton University and supported by N.S.F. under Grant GK-1439 and the U.S.A.F. under Grant 1333-67

We can also pick the weighing matrix Ψ tobe a function of s so as to take into consideration the importance of the system sensitivity at the various operating frequencies.

Our design objective will be the minimization of the family $\{I(\underline{x})\}_{\underline{x}\varepsilon X}$ where X is the parameter space. This can be accomplished by the minimization of $\max_{\underline{x}\varepsilon X} I(\underline{x})$. This " minimax" approach is often too pessimistic. In addition, the solution of the minimax problem depends upon the values of $\underline{x}$ on the boundary of X and this boundary is rarely known with certainty. A more realistic approach would be to weigh the values that F(s) assumes, according to their frequency of appearance and/or undesirability.Thus, a probability density $p(\underline{x})$ could be implicitly defined on X, and the design objective would be the minimization of the weighted average of $I(\underline{x})$:

$$J=EI(\underline{x}) = \frac{1}{2\pi j} E \int_{-j\infty}^{+j\infty} [\underline{y}-\underline{y}_o]_* \Psi [\underline{y}-\underline{y}_o] ds \tag{1}$$

If input variations are also present, the expectation in (1) could refer to both Plant parameter and input disturbances.

Another difference between the univariable and multivariable sensitivity problems concerns the degree of freedom present in each problem. For the unconstrained single-input, single-output case J will be minimized if

$$E\, \underline{y}(s) = \underline{y}_o(s) \tag{2}$$

This relation determines the scalar filter H(s) completely and no further minimization of J is possible. In the n-input, n-output case, however, we have n^2 elements of H(s) to choose while (2) represents only n equations. Thus (2) can be taken as a mere constraint on the minimization of J. We can show that under the above formulation feedback control is generally superior to open-loop control in the multivariable case. For an intuitive argument consider the open-loop configuration of Fig.2. There $\underline{y} = FG\underline{u}$ and the only control variables that influence J will be the n components of the vector $G\underline{u}$. Thus we only have freedom enough to ensure (at best) the validity of relation (2), while with a feedback configuration we can proceed considerably beyond that.

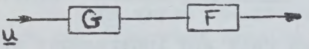


FIG. 2

These observations lead us to formulate our problem as follows:

We shall consider the feedback system of Fig. 1 and we shall seek a controller H which minimizes the performance index J given by (1) under constraint (2) and the additional requirement that the nominal composite transfer matrix of the system be stable.

The necessary analysis tools are provided in the Appendix where a general constrained multivariable minimization problem is solved.

2. PRELIMINARY ANALYSIS

The composite transfer matrix of the system in Fig. 1 is $(I_n+FH)^{-1}F$ where I_n is the n-dimensional identity matrix. Let $F_o(s)$ denote the expectation of F(s):

$$F_o(s) = E\, F(s) \tag{3}$$

Define

$$G(s) \triangleq (I_n+F_oH)^{-1} \tag{4}$$

To simplify the analysis we shall assume that the variance of each element of F is very small compared to 1 and thus we shall use the approximation:

$$(I_n+FH)^{-1}F = (I_n+G(F-F_o)H)^{-1}GF \cong GF-G(F-F_o)HGF_o \quad .$$ Then

$$\underline{y} = (I_n+FH)^{-1}F\,\underline{u} \cong GF\underline{u} - G(F-F_o)HGF_o\underline{u} \tag{5}$$

From (5) $E\,\underline{y} = GF_o\underline{u}$ and constraint (2) is equivalent to:

$$\underline{y}_o = GF_o\underline{u} \tag{6}$$

From (5) and (6):

$$\underline{y} - \underline{y}_o = G(F-F_o)\,(I_r-HGF_o)\underline{u} \tag{7}$$

If the number r of the inputs is larger than the number n of the outputs, we can concentrate on any n inputs that control the plant. Also the minimization of J can be seen to involve n inputs at the most. Thus in the sequel we shall restrict ourselves to the case where $r \leq n$. When $r=n$, F_o and H are square matrices. Assuming the normal rank of F_o to be n: $H = F_o^{-1}(G^{-1} - I_n)$. The difference $\underline{y}-\underline{y}_o$ of (7) becomes in this case

$$\underline{y}-\underline{y}_o = G(F-F_o)\underline{u}_o \tag{8}$$

where $\underline{u}_o = F_o^{-1}y_o$. In (8) the only unknown matrix is G.

When $r<n$ we assume that the normal rank of F_o is r. From $G^{-1} = I_n+F_oH$ we obtain

$$GF_o = F_o[I_r+HF_o]^{-1} \tag{9}$$

also

$$\underline{y}_o = GF_o\underline{u} = F_o[I_r+HF_o]^{-1}\underline{u} \tag{10}$$

From (10) we see that there always exists a vector $\underline{u}_o = (I_r+HF_o)^{-1}\underline{u}$ such that

$$\underline{y}_o = F_o\underline{u}_o \tag{11}$$

In other words if the desired output y_o is to be at all attainable from the system under consideration, it must be of the form (11). We can, therefore, assume that (11) always has a solution for $\underline{u}_o$. If $r=n$ then, of course, $\underline{u}_o = F_o^{-1}y_o$. Now $(I_r+HF_o)\,(I_r-HGF_o) = I_r-HGF_o+HF_o-HF_oHGF_o$. Since $F_oHG = I-G$, the right side equals I_r. Thus $(I_r+HF_o)^{-1} = I_r-HGF_o$ and

$$(I_r-HGF_o)\underline{u} = (I_r+HF_o)^{-1}\underline{u} = \underline{u}_o \tag{12}$$

From (7) and (12) we obtain $\underline{y}-\underline{y}_o = G(F-F_o)\underline{u}_o$ which is identical to relation (8). Clearly for the case $r<\overline{n}$, H must satisfy an additional constraint:

$$(I_r + HF_o)^{-1}\underline{u} = \underline{u}_o \tag{13}$$

But we can see easily that the constraint already imposed $GF_o\underline{u} = \underline{y}_o$ is equivalent to $F_o[(I_r+HF_o)^{-1}\underline{u}-\underline{u}_o] = 0$ and thus equivalent to (13) since the normal rank of F_o has been assumed to be r.

To summarize, we have shown that if $r \leq n$ the equation $\underline{y}_o = F_o\underline{u}_o$ has a solution for $\underline{u}_o$ and $\underline{y}-\underline{y}_o = G(F-F_o)\underline{u}_o$. Thus the index of performance J will become

$$J = \frac{1}{2\pi j}\,E\int_{-j\infty}^{+j\infty} \underline{u}_{o*}(F-F_o)_*G_*\Psi\,G(F-F_o)\underline{u}_o\,ds\ .$$

We have assumed that the matrix $\Psi(s)$ is para-Hermitian and nonnegative definite for all imaginary s. Then Ψ [7] can be factored into A_*A where

$A(s)$ and its inverse $A^{-1}(s)$ are analytic in the Right Half-Plane (RHP). The index J will become

$$J = \frac{1}{2\pi j} E\int_{-j\infty}^{+j\infty} \underline{u}_{o*}(F-F_o)_* G_* A_* A G (F-F_o)\underline{u}_o \, ds =$$

$$= \frac{1}{2\pi j} \text{trace } \{E\int_{-j\infty}^{+j\infty} AG(F-F_o)\underline{u}_o\underline{u}_{o*}(F-F_o)_* G_* A_* \, ds\}$$

Define

$$R(s) \triangleq E\,\{(F-F_o)\underline{u}_o\underline{u}_{o*}(F-F_o)_*\} \tag{14}$$

R is independent of the unknown controller H and

$$J = \frac{1}{2\pi j} \text{trace} \int_{-j\infty}^{+j\infty} AGRG_*A_* \, ds$$

The first of the constraints to be considered is $GF_o\underline{u} = \underline{y}_o$ which is equivalent to $AGF_o\underline{u} = A\underline{y}_o$, and the second is the constraint that GF_o be stable, i.e., analytic in RHP. But, since $A(s)$ and $A^{-1}(s)$ are analytic in RHP, GF_o will be stable if and only if AGF_o is stable. Also $A\underline{y}_o$ is analytic in RHP. Therefore, with no loss of generality, we can take $A = I_n$ in the sequel. The solution for $A \neq I_n$ can be obtained if AG is substituted for G and $A\underline{y}_o$ is substituted for $\underline{y}_o$ in the solution for the case $A = I_n$.

Thus the problem reduces to the determination of a matrix G(s) minimizing J:

$$J = \frac{1}{2\pi j} \text{trace} \int_{-j\infty}^{+j\infty} GRG_* ds \tag{15}$$

under the constraints

a) $GF_o\underline{u} = \underline{y}_o$ and

b) GF_o is analytic in RHP, if stability is required.

We shall accomplish the minimization in two steps. Let us first define:

$$Y(s) \triangleq [I_r + HF_o]^{-1} \tag{16}$$

We shall determine the H(s) that minimizes J while keeping Y(s) constant. The change in H will be constrained by relation (16) or, equivalently, by $HF_o = Y^{-1} - I_r$. From (9) $GF_o = F_oY$. Therefore, by keeping Y constant we are in effect keeping GF_o constant. Since the two constraints a) and b) above only refer to GF_o, they will not concern us in this minimization step. The application of standard variational techniques leads to the following expression for the optimum H:

$$H_o = Y^{-1}(F_{o*}F_o)^{-1}F_{o*} - (F_{o*}R^{-1}F_o)^{-1}F_{o*}R^{-1} \tag{17}$$

The corresponding G is

$$G = I_n - F_o(F_{o*}F_o)^{-1}F_{o*} + F_oY(F_{o*}R^{-1}F_o)^{-1}F_{o*}R^{-1} \tag{18}$$

and the corresponding J:

$$J = \frac{1}{2\pi j} \text{trace} \int_{-j\infty}^{+j\infty} GRG_* \, ds = J_1 + J_2 \text{ where}$$

$$J_1 = \frac{1}{2\pi j} \text{trace} \int_{-j\infty}^{+j\infty} [I_n - F_o(F_{o*}F_o)^{-1}F_{o*}]R[I_n - F_o(F_{o*}F_o)^{-1}F_{o*}] \, ds \tag{19}$$

$$J_2 = \frac{1}{2\pi j} \text{trace} \int_{-j\infty}^{+j\infty} GF_o(F_{o*}R^{-1}F_o)^{-1}F_{o*}G_* \, ds \tag{20}$$

The first component J_1 of J does not depend on G and is zero when r=n and $F_o(s)$ is full normal rank. It only remains to minimize J_2 with respect to G. This minimization depends on whether or not stability is required and is discussed in the next two Sections.

3. THE UNCONSTRAINED SYSTEM OPTIMIZATION

Define $X(s) \triangleq G(s)F_o(s) = F_o(s)Y(s)$. It is desired to minimize the J_2 given by

$$J_2 = \frac{1}{2\pi j}\,\mathrm{trace}\int_{-j\infty}^{+j\infty} X(F_{o*}R^{-1}F_o)^{-1}X_*\,ds$$

under the constraint $X\underline{u} = \underline{y}_o$ and the requirement that a rxr matrix Y(s) exist such that $X(s) = F_o(s)Y(s)$.

This is the special case of the problem of the Appendix, with $\Theta = 0$ and the realizability constraint removed. Its solution is easily seen to be:

$$X_{opt.}(s) = F_o(s)Y_{opt.}(s) = \frac{\underline{y}_o\underline{u}_*F_{o*}R^{-1}F_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}} \tag{21}$$

And the minimum J_2

$$J_{2m} = \frac{1}{2\pi j}\int_{-j\infty}^{+j\infty}\frac{\underline{y}_{o*}\underline{y}_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}}\,ds \tag{22}$$

From (21)

$$Y_{opt.} = \frac{\underline{u}_o\underline{u}_*F_{o*}R^{-1}F_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}} \tag{23}$$

The optimum H must be found from (17) and (23).

Two serious disadvantages of this solution should be noted here. First, the $Y_{opt.}$ given by (23) is singular, and hence its inverse, needed for the determination of H does not exist. One way to overcome this difficulty would be to search for a suboptimal solution $Y = Y_{opt.} + K(s)$ with $K\underline{u} = \underline{0}$. Then $GF_o\underline{u} = F_oY\underline{u} = \underline{y}_o$. The index J_2 will be augmented by

$$J_2(\varepsilon)-J_{2m} = \frac{\varepsilon^2}{2\pi j}\,\mathrm{trace}\int_{-j\infty}^{+j\infty} F_oK(F_{o*}R^{-1}F_o)^{-1}K_*F_{o*}\,ds\ .$$

and this difference can be made arbitrarily small.

The second disadvantage of the solution given by (23) is that the resulting nominal composite transfer matrix F_oY_{opt} is not stable in general. Thus, while the system is certainly stable for inputs close to the reference input $\underline{u}(s)$ if F(s) has the value $F_o(s)$, the variations of parameters in F may cause instability. This disadvantage can be overcome either by a careful selection of the matrix K mentioned above or by imposing a stability constraint on YF_o as to be discussed in the next Section.

4. THE CONSTRAINED SYSTEM OPTIMIZATION

As in the previous Section we define $X(s) = G(s)F_o(s) = F_o(s)Y(s)$. Then we need to determine X so that

$$J_2 = \frac{1}{2\pi j}\,\mathrm{trace}\int_{-j\infty}^{+j\infty} X(F_{o*}R^{-1}F_o)^{-1}X_*\,ds$$

is minimized under the constraint $X\underline{u} = \underline{y}_o$ and the requirements that X(s) be analytic in RHP and that there exist a rxr matrix Y(s) such that $X(s) = F_o(s)Y(s)$.

Again this is the special case of the problem examined in the Appendix with $\Theta = 0$, m = r. Its solution is

$$X_o(s) = F_oY_o = G_oF_o = \frac{\underline{y}_o\underline{u}_*F_{o*}R^{-1}F_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}_*} + V\left[V^{-1}\frac{\underline{y}_o\underline{u}_*F_{o*}R^{-1}F_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}}W^{-1}\right]_- W \tag{24}$$

where the (r-1)xr matrix W is analytic in RHP and satisfies

$$W_*W = F_{o*}R^{-1}F_o - \frac{F_{o*}R^{-1}F_o\underline{u}\underline{u}_*F_{o*}R^{-1}F_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}} \tag{25}$$

and W^{-1} is any right inverse of W that is analytic in RHP.

Similarly, the nxr matrix V is analytic in RHP and satisfies

$$V\,V_* = F_o(F_{o*}F_o)^{-1}F_{o*} \tag{26}$$

and V^{-1} is any left inverse of V that is analytic in RHP.

From (24) the optimum Y(s) is given by

$$Y_o = \frac{\underline{u}_o\underline{u}_*F_{o*}R^{-1}F_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}} + (F_{o*}F_o)^{-1}F_{o*}V\left[V^{-1}\frac{\underline{y}_o\underline{u}_*F_{o*}R^{-1}F_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}}W^{-1}\right]_- W \tag{27}$$

G_o can be determined from $G_oF_o = F_oY_o$. And the optimum H is found from (17) and (27), or directly from $I_n+F_oH = G_o^{-1}$.

The minimum J_2 is calculated to be

$$J_{2min} = \frac{1}{2\pi j}\int_{-j\infty}^{+j\infty}\frac{\underline{y}_{o*}\underline{y}_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}}\,ds$$

$$+\frac{1}{2\pi j}\,\text{trace}\int_{-j\infty}^{+j\infty}\left[V^{-1}\frac{\underline{y}_o\underline{u}_*F_{o*}R^{-1}F_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}}W^{-1}\right]_-\left[V^{-1}\frac{\underline{y}_o\underline{u}_*F_{o*}R^{-1}F_o}{\underline{u}_*F_{o*}R^{-1}F_o\underline{u}}W^{-1}\right]_{-*}ds \tag{28}$$

Since the first term is, according to (22) the minimum J_2 for unconstrained GF_o , the second term represents the increase in J_{2min} due to the requirement that GF_o be stable.

Finally, the minimum of the overall J is $J_1 + J_{2min}$ where J_1 is given by (19) and J_{2min} by (28).

5. REALIZATION OF THE CONTROLLER

While the controller H(s) defined from (17) and (27) is not, in general, physically realizable as an open loop structure, it can always be realized in a simple closed loop configuration. For example, H is easily seen to be equal to the composite transfer matrix of the realizable structure in Fig. 3 where L_1 , L_2 are any constant nxr matrices and U_1 , U_2 any constant rxn matrices satisfying

$$U_1L_1 = U_2L_2 = I_r \quad \text{and}$$

$$Z_1 = GF_o - L_1$$

$$Z_2 = F_o - L_2 \ .$$

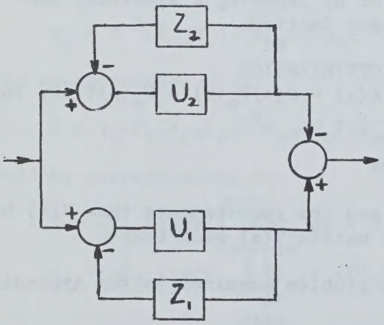


FIG. 3

APPENDIX

A Multivariable Minimization Problem

In this Appendix we state a general minimization problem and determine its solution. The results to be obtained can be applied to the sensitivity problem formulated in Section 1, as well as to other mean square minimization problems.

Let F(s) be a nxm matrix with normal rank m where $r \leq m \leq n$. Define $T(s) = I_n - F(F_*F)^{-1}F_*$ where I_n is the n-dimensional Identity matrix and F_* is a shorthand notation for $F^T(-s)$. Let $\underline{y}(s)$ and $\underline{u}(s)$ be n- and r-dimensional column vectors that are analytic and bounded in RHP. Let $\underline{y}(s) = F(s)\underline{v}(s)$ where $\underline{v}(s)$ is an m-dimensional column vector.

Consider the problem of determining a nxr matrix X(s) to minimize J, where

$$J = \frac{1}{2\pi j} \text{ trace} \int_{-j\infty}^{+j\infty} (X\,X_* - 2\Theta X_*)ds \tag{29}$$

under the constraints

a) X(s) is analytic in RHP and bounded.

b) $X(s)\underline{u}(s) = \underline{y}(s)$

c) $T(s)X(s) = 0$

We shall show that if the solution $X_o(s)$ to this problem exists, then

(i) $$X_o(s) = \frac{\underline{y}\underline{u}_*}{\underline{u}_*\underline{u}} + V\left[V_*\Theta W_*\right]^+ W - V\left[V^{-1}\,\frac{\underline{y}\underline{u}_*}{\underline{u}_*\underline{u}}\,W^{-1}\right]^- W \tag{30}$$

where the superscript + indicates that term of additive decomposition of a matrix, which is analytic in RHP and tends to zero as s→∞. The superscript - indicates the corresponding term which is analytic in LHP. The (r-1)xr matrix W is defined by the spectral factorization:

$$W_*W = I_r - \frac{\underline{u}\,\underline{u}_*}{\underline{u}_*\underline{u}}$$

and the requirement that W(s) and its right inverse $W^{-1}(s)$ be analytic in RHP. Similarly, the nxm matrix V(s) is defined by

$$V\,V_* = F(F_*F)^{-1}F_* = I_n - T$$

and the requirement that V(s) and its left inverse $V^{-1}(s)$ be analytic in RHP.

(ii) Although the matrices W, V, W^{-1}, V^{-1} defined above are not unique, the matrix $X_o(s)$ obtained from (30) is unique.

(iii) The minimum J is given by

$$J_{min} = \frac{1}{2\pi j}\int_{-j\infty}^{+j\infty}\left(\frac{\underline{y}_*\underline{y}}{\underline{u}_*\underline{u}} - 2\,\frac{\underline{u}_*\Theta_*\underline{y}}{\underline{u}_*\underline{u}}\right)ds$$

$$+ \frac{1}{2\pi j} \text{ trace}\int_{-j\infty}^{+j\infty}\left([V_*\Theta W_*]^+\;[V_*\Theta W_*]^+_* - 2V[V_*\Theta W_*]^+\Theta_*\right)ds$$

$$+ \frac{1}{2\pi j} \text{ trace}\int_{j\infty}^{+j\infty}\left(\left[V^{-1}\,\frac{\underline{y}\underline{u}_*}{\underline{u}_*\underline{u}}\,W^{-1}\right]^-\left[V^{-1}\,\frac{\underline{y}\underline{u}_*}{\underline{u}_*\underline{u}}\,W^{-1}\right]^-_* - 2V\left[V^{-1}\,\frac{\underline{y}\underline{u}_*}{\underline{u}_*\underline{u}}\,W^{-1}\right]^-\Theta_*\right)ds \tag{31}$$

In order to avoid complication of the proof we shall make two, otherwise unnecessary, additional assumptions. We shall assume that all vectors and matrices considered are real, rational functions of s and that $T(s)$, $\Theta(s)$ and $\frac{\underline{u}(s)\underline{u}_*(s)}{\underline{u}_*(s)\underline{u}(s)}$ are bounded in the s-plane.

Proof of Assertion (i). Through application of standard variational techniques we can easily conclude that if X minimizes J under the constraints a), b), c) above, then

$$X = \frac{\underline{y}\underline{u}_*}{\underline{u}_*\underline{u}} + (I_n-T)\,\Theta\left(I_r - \frac{\underline{u}\underline{u}_*}{\underline{u}_*\underline{u}}\right) + (I_n-T)B^-\left(I_r - \frac{\underline{u}\underline{u}_*}{\underline{u}_*\underline{u}}\right) \tag{32}$$

where the nxr matrix B^- is analytic in the Left Half-plane (LHP) but otherwise unspecified. To determine B^- we note first that the rxr matrix $I_r - \frac{\underline{u}\underline{u}_*}{\underline{u}_*\underline{u}}$ is para-Hermitian, bounded, nonnegative definite for all imaginary s and has normal rank r-1. Then [7] there exists a (r-1)xr matrix W(s) analytic and bounded in the RHP, with at least one right inverse W^{-1} analytic in the RHP and such that

$$I_r - \frac{\underline{u}\underline{u}_*}{\underline{u}_*\underline{u}} = W_*W \tag{33}$$

In addition [7] any other matrix W' with the same properties is related to W by

$$W' = Q\,W \tag{34}$$

where Q is a constant, square, unitary matrix

$$Q^TQ = I_{r-1} \tag{35}$$

In addition we can easily prove that all such matrices W satisfy:

$$W\underline{u} = \underline{0} \tag{36}$$

$$WW_* = I_{r-1} \tag{37}$$

Similarly, the matrix $I_n - T = F(F_*F)^{-1}F_*$ can be factored into:

$$I_n - T = VV_* \tag{38}$$

where nxm matrix V(s) is analytic and bounded in RHP and has at least one left inverse V^{-1} analytic in RHP. Any other matrix V' with the same properties is related to V by

$$V' = VR \tag{39}$$

where R is a constant, square, unitary matrix

$$RR^T = I_m \tag{40}$$

In addition we easily prove that

$$TV = 0 \tag{41}$$

$$T\underline{y} = \underline{0} \tag{42}$$

$$V_*V = I_m \tag{43}$$

Now, from (32), (33), (37) we have

Now, from (32), (33), (37) we have

$$V^{-1} X \, W^{-1} = V^{-1} \frac{\underline{y} \underline{u}_*}{\underline{u}_* \underline{u}} W^{-1} + V_* \Theta \, W_* + V_* B^- W_*$$

By equating the analytic in LHP partial franctions in the above equation, we obtain

$$V_* B^- W_* = -\left[V^{-1} \frac{\underline{y} \underline{u}_*}{\underline{u}_* \underline{u}} W^{-1} + V_* \Theta \, W_* \right]^- - C(s)$$

from which

$$X = \frac{\underline{y} \underline{u}_*}{\underline{u}_* \underline{u}} - V \left[V^{-1} \frac{\underline{y} \underline{u}_*}{\underline{u}_* \underline{u}} W^{-1} \right]^- W + V \left[V_* \Theta \, W_* \right]^+ W - VCW \tag{44}$$

where C(s) is an unknown mx(r-1) matrix whose elements are real polynomials in s. Thus, if an optimum X exists, it will be of the form (44). To conclude the proof we note that if we substitute (44) into (29) the resulting J will not, in general, be finite unless C = 0. Therefore if an optimum X exists, it is given by (30).

Proof of Assertion (ii). The fact that the X_o given by (30) is independent of the particular selection for W,V is a trivial consequence of (34), (35), (39), (40). Also, if V_1^{-1}, V_2^{-1} are any two left inverses of V: $(V_1^{-1} - V_2^{-1})V = 0$. From (38) : $(V_1^{-1} - V_2^{-1})F = 0$ and hence $(V_1^{-1} - V_2^{-1}) \, \underline{y} = \underline{0}$. Thus

$$X_o(V_1^{-1}) - X_o(V_2^{-1}) = V \left[(V_2^{-1} - V_1^{-1}) \frac{\underline{y} \underline{u}_*}{\underline{u}_* \underline{u}} W^{-1} \right]^- W = 0$$

Finally if W_1^{-1}, W_2^{-1} are two right inverses of W, both analytic in RHP, $W(W_1^{-1} - W_2^{-1}) = 0$. But the normal rank of W is r-1. Hence the only solution of $W\underline{x} = \underline{0}$ is $\underline{x} = \underline{u}$. Therefore $W_1^{-1} - W_2^{-1} = \underline{u}(s)\underline{\sigma}^T(s)$. Since W_1^{-1}, W_2^{-1} are analytic in RHP, $\underline{\sigma}(s)$ is also analytic in RHP. Thus

$$X_o(W_1^{-1}) - X_o(W_2^{-1}) = - V[V^{-1} \underline{y} \underline{\sigma}^T]^- \, W = 0$$

Because both V^{-1} and $\underline{y}\underline{\sigma}^T$ are analytic in RHP.

The *Proof of Assertion (iii)* is immediate from relations (30), (36), (37), (41), (43).

REFERENCES

[1] See for example: John B. Thomas: "On Systems with Unknown or Uncertain Parameters." *Proc. 6th Ann. Allerton Conf. on Circuit and System Theory*, October 1968, pp. 634-645.

[2] B. Liu and M. E. Van Valkenburg: "A Probabilistic Approach to the Parameter Sensitivity Problem." *Proc. 5th Ann. Allerton Conf. on Circuit and System Theory*, October 1967, pp. 701-710.

[3] J. B. Cruz, Jr. and W. R. Perkins: "A New Approach to the Sensitivity Problem in Multivariable Feedback System Design." IEEE Trans. on Automatic Control, vol. AC-9, pp. 216-223, July 1964.

[4] R. L. Gonzales: "Synthesis for Minimum Sensitivity Under Worst Case Conditions." Proc. 4th Ann. Allerton Conf. on Circuit and System Theory, October 1966, 1966, pp. 527-537.

[5] William R. Perkins, Jose B. Cruz and Richard L. Gonzales: "Design of Minimum Sensitivity Systems." IEEE Trans. on Automatic Control, vol. AC-13, pp. 159-168, April 1968.

[6] I. M. Horowitz: Synthesis of Feedback Systems, Academic Press, New York; 1963.

[7] D. C. Youla: "On the Factorization of Rational Matrices." IRE Trans. on Information Theory, vol. IT-7, pp. 172-189, July 1961.

[8] Norbert Wiener: Extrapolation, Interpolation and Smoothing of Stationary Time Series. MIT Press, Cambridge, Mass.; 1949.

TIME OPTIMAL LINEAR TIME-LAG SYSTEMS WITH MULTIPLE CONSTRAINTS ON THE CONTROL FUNCTION

DONG HAK CHYUNG
Department of Electrical Engineering
The University of Iowa
Iowa City, Iowa 52240

ABSTRACT

The linear time-lag system

$$\dot{x}(t) = A_0\, x(t) + A_1\, x(t-h) + B\, u(t)$$

with the constraints $|u(t)| \leq 1$, $|\dot{u}(t)| \leq 1$ on the scalar control function u(t) is considered. The problem is to find an optimal control which steers the response of the system from a given initial state $x(t) = \phi(t)$ on $[-h,0]$ to the origin $x = 0$ in minimum time. Necessary conditions for an optimal control are obtained in the paper.

Consider a linear time-lag system with delay $h > 0$ which is modeled by the vector differential-difference equation

$$\dot{x}(t) = A_0\, x(t) + A_1\, x(t-h) + B\, u(t).$$

For the sake of simplicity we assume that A_0, A_1 and B are constant matrices and the control function u(t) is a scalar. The admissible control functions are absolutely continuous functions u(t) such that $|u(t)| \leq 1$ and $|\dot{u}(t)| \leq 1$, that is, the velocity, as well as the magnitude, of the control function is limited. Thus the controller changes continuously with a limit on the maximum rate of change. The rudder of an airplane is an example of such controllers. This is, of course, a special case of bounded phase coordinate problems.

The problem is to find an admissible control which steers the response of the system from initial states $x(t) = \phi(t)$, $-h \leq t \leq 0$, $u(0) = u_0$ to the origin $x = 0$ in minimum time T with $u(T) = u_1$. Here $\phi(t)$, u_0 and u_1 are given. By a similar argument as in Chapter 5 of [1] the following results can be obtained.

Let $u^*(t)$ with response $x^*(t)$ on $[0,T]$ be a time optimal control with the minimum time T. Then there exists a nontrivial continuous vector solution $p(t)$ of the equation

$$\dot{p}(t) = \begin{cases} -p(t)A_0 - p(t+h)A_1 & , \quad 0 \leq t \leq T-h \\ -p(t)A_0 & , \quad T-h \leq t \leq T \end{cases}$$

such that

$$\int_0^T p(t)B\, u^*(t)\, dt = \max \int_0^T p(t)\, Bu(t)\, dt$$

for all admissible controls with $u(0) = u_0$, $u(T) = u_1$. Furthermore if the system is normal on $[0,T]$ then either $|u^*(t)| = 1$ or $|\dot{u}^*(t)| = 1$ almost everywhere on $[0,T]$. The system is called normal if no component of $qY(t)B$ is identically zero on any positive subinterval of $[0,T]$ for all nonzero vectors q, where $Y(t)$ is the continuous matrix solution of

$$\dot{Y}(t) = \begin{cases} -Y(t)A_0 - Y(t+h)A_1 & , \quad 0 \leq t \leq T-h \\ -Y(t)A_0 & , \quad T-h \leq t \leq T \end{cases}$$

with $Y(T) = I$, the identity matrix.

REFERENCES

[1] E.B. Lee and L. Markus, "Foundations of Optimal Control Theory," John Wiley, New York, 1967.

[2] L.S. Pontryagin, et al., "The Mathematical Theory of Optimal Processes," Interscience, New York, 1962.

[3] S.S.L. Chang, "Minimum Time Control with Multiple Saturation Limits," IEEE T-AC, Vol. AC-8, No. 1, 1963.

[4] W.W. Schmaedeke and D.L. Russell, "Time Optimal Control with Amplitude and Rate Limited Controls," SIAM Journal on Control, Vol. 2, No. 3, 1965.

[5] D.H. Chyung and E.B. Lee, "Control of Linear Time Delay Systems with Essentially Linear Cost Functionals," in Differential Equations and Dynamical Systems, J.P. LaSalle, ed., Academic Press, New York, 1967.

[6] D.H. Chyung and E.B. Lee, "Linear Optimal Systems with Time Dealys," SIAM J. on Control, Vol. 4, No. 3, 1966.

COMPUTATION OF GENERALIZED INVERSE AND MATRIX RANK

NARESH N. GUPTA
Wright State University
Dayton, Ohio

ABSTRACT

In this paper, we develop a method for computing the generalized inverse A^+ and the associated projections AA^+ and A^+A for any matrix A. The method does not use the knowledge of the dimension of range space of A. It is also shown that the trace of a sequence of approximations to AA^+ (or A^+A) converges to the rank of A.

Many communication and control problems reduce to the solution of the general normal equations $X'V^*X\beta = X'V^*Y$...(1) where Y is an n×1 vector of observations, X is an n×p matrix of known real constants, β is a p×1 vector of unknown parameters and σ^2V is a known covariance matrice of errors. In general V is nonnegative and possibly singular. In such a case the solution(1) of (1) involves the generalized inverse of V. Ben Israel and Cohen(2) gave an iterative method with convergence of order two for computing generalized inverse. In this paper we generalize that method so as to have convergence of any desired finite order.

The main results are the Theorems(3) 1, 2, and 3. Theorems 1 and 2, respectively, show the convergence of a sequence of approximations to the generalized inverse and the associated projections. Theorem 3 shows that the trace of a sequence of approximations to AA^+ (or A^+A) converges to the rank of A.

Theorem 1

The sequence of transformations defined by $S_{k+1} = S_k[I+(I-AS_k) + (I-AS_k)^2 + \ldots + (I-AS_k)^n]$ with $S_o = \alpha A^*$ and $0<\alpha<2/\lambda\max$ converges to the generalized inverse A^+ of A. $\lambda\max$ is the maximum eigenvalue of AA^* or (A^*A).

Theorem 2

a) The sequence of transformations defined by $AS_{k+1} = AS_k[I+(I-AS_k) + (I-AS_k)^2 + \ldots + (I-AS_k)^n]$ with S_o and α as defined in Theorem 1 converges to the orthogonal projection AA^+.

b) The sequence of transformations defined by $S_{k+1}A = S_k[I+(I-AS_k) + (I-AS_k)^2 + \ldots + (I-AS_k)^n]A$ with S_o and α as in (a) converges to the orthogonal projection A^+A.

Theorem 3

The trace of $X_k(=AS_k)$ or $(=S_kA)$ is monotone increasing function of k, k=1, 2, ..., converging to the rank of A.

Remark: In using Theorems 1, 2, and 3 it is not necessary to compute λ_{max}. Writing $AA^* = (d_{ij})(i,j=1,\ldots,n)$ we conclude from Gershgorin theorem that,

$$\lambda_{max} \leq \max_i \left[\sum_{j=1}^{n} |d_{ij}| \right].$$

Using Theorem 3 to determine the rank of any matrix A, we need only to compute the sequence of approximations to the trace of AA^+ or A^+A whichever is of lower order.

REFERENCES

1. G. Zyskind and F.B. Martin, "On Best Linear Estimation and a General Gauss-Markov Theorem in Linear Models with Arbitrary Nonnegative Convariance Structure," J. Siam Applied Math, 6(1969), pp. 1190-1202.

2. A. Ben-Israel and D. Cohen, "On Iterative Computation of Generalized Inverses and Associated Projections," J. Siam, Numerical Analysis, 3(1969), pp. 410-419.

3. N.N. Gupta, "The Determination of the Order of a Linear Differential System with Noisy Observations," Ph.D. Thesis, Polytechnic Institute of Brooklyn, 1970.

A DIRECT SEARCH FOR OPTIMAL CONTROL OF DISTRIBUTED SYSTEMS

Albert J. Faulstich, Jr. and Hiralal C. Khatri
Mechanical Engineer
Naval Ordnance Laboratory
White Oak, Md. 20910

Associate Professor of
Mechanical Engineering
Catholic University of
America
Washington, D. C. 20017

ABSTRACT

Digital computer oriented algorithms for direct search of optimal control of distributed parameter systems are given. The optimal controller minimizes an integral (single or multiple) type index of performance subject to system integral (single or multiple) equation constraints. Side constraints on the system variables are incorporated using penalty functions. The algorithms are derived using first variations. Results of digital computer simulations for a lumped parameter problem, a heat diffusion problem, and a beam in transverse vibrations are given.

INTRODUCTION

Investigation of the optimal control of systems described by ordinary differential equations (lumped systems) has been extensive. The results are well known [1,2]*. During this decade optimum control of distributed systems (systems described by partial differential equations or equivalent integral equations) has been considered by many authors. Some fundamental contributions in this area are credited to Butkovskii [3,4,5,6,7], Lerner [3], and Lurie [8,9,10]. Advances have been made in constructing algorithms for distributed systems using techniques such as dynamic programming [10,11] and variation techniques for partial differential equations. J. L. Lions [12] has authored a text recently covering various aspects of the subject. The ASME has published a collection of papers [13] that surveys the field and covers recent developments. Each of these papers has an extensive bibliography.

Many distributed parameter systems may be described by either partial differential equations and their associated boundary conditions or equivalent integral equations. If differential equations and boundary conditions are used for a system model, the numerical solution for the optimal controller requires excessive digital computer memory and operation. Memory and operation time are needed to solve differential equations that are part of the algorithm being used. Whereas, if an integral model can be obtained, computer memory and operation may be reduced somewhat due to a fewer number of calculations. It should be noted that it is not always easy to secure an integral equation model.

Algorithms for the optimization of distributed parameter systems described by integral equations are presented here. These algorithms are derived using the first order relaxation technique. The method used is an extension of the technique in Merriam [2] for ordinary differential equation contraints. The calculations were performed on an IBM 7090.

The method is derived and applied to the following types of examples:

a. Lumped Parameter Problem

---

*Numbers in brackets indicate references at the end of the paper.

b. Distributed System with Non-Distributed Input

c. Distributed System with Distributed Input

Example (a) is included so that the results using the algorithm discussed in this paper could be compared with a known exact closed form solution. The comparison is excellent. Example (b) is a heat diffusion problem. It includes side constraints in addition to the index of performance. These side constraints are introduced as penalty functions. Some of the results of example (b) are compared with infinite series type solutions found in the literature. Again, the comparison is excellent. The problem selected for example (c) is that of a vibrating beam with a distributed load.

PROBLEM STATEMENT

We shall consider the optimal control of distributed systems with a nondistributed input and then a distributed input.

a. Distributed system with nondistributed input: Let the system be described by a linear integral equation

$$v(x,t) = \int_{t_o}^{t} W(x,t,\tau)\, u(\tau)\, d\tau + h(x,t) \tag{1}$$

where v, the controlled variable, is n-dimensional; u, the control variable, is m-dimensional; x is p-dimensional; $x\varepsilon\Omega$; t denotes time. The function W is a nxm matrix. h(x,t) is an n-dimensional vector representing contributions due to initial conditions. For this system, we want to determine the controller which minimizes the index of performance

$$I = \int_{\Omega} dx \int_{t_0}^{t_f} dt\; F(v(x,t),u(t),x,t) \tag{2}$$

$$\text{or } I = \int_{t_o}^{t_f} dt\; F(v(x_2,t),u(t),t)\; ;\; x_2\varepsilon\Omega \tag{3}$$

subject to system equation (1). The scalar function F includes penalty functions due to any side constraints.

b. Distributed parameter system with distributed input:

Let the system be described by a linear integral equation of the type

$$v(x,t) = \int_{\Omega} d\gamma \int_{t_o}^{t} d\tau\; W(x,\tau,t,\tau)\; u(\gamma,\tau) + h(x,t) \tag{4}$$

where the vectors and functions are as defined before.
Determine the controller which minimizes

$$I = \int_{\Omega} dx \int_{t_o}^{tf} dt\; F(v(x,t),u(x,t),x,t) \tag{5}$$

subject to system equation (4).

RECURSIVE ALGORITHMS

The algorithms for these problems are:

a. Nondistributed Input

For a system defined by equation (1) and index of performance (3), a recursive algorithm for finding the optimal controller is given by the incremental change in the controller

$$\delta u(t) = -\varepsilon \operatorname{sgn} \left[\int_t^{t_f} W^T(\tau,t) \frac{\partial F(\tau)}{\partial v(\tau)} d\tau + \frac{\partial F(t)}{\partial u(t)} \right] \tag{6}$$

where ε is the increment step and the sgn function is defined as

$$\operatorname{sgn}(x) = \begin{cases} 1 , & x > 0 \\ 0 , & x = 0 \\ -1 , & x < 0; \end{cases} \tag{7}$$

W^T is the transpose of W. The value of the increment step, ε, is chosen to be a constant. There is no unique determination of ε. The iteration scheme for a scalar example is shown in Figure 1. New selections for u are computed

$$u_{i+1} = u_i + \delta u \tag{8}$$

The change in the controlled variable, δv, is computed from equation (1) by replacing u by δu. The new values for v are computed

$$v_{i+1} = v_i + \delta v \tag{9}$$

Using u_{i+1} and v_{i+1}, I_{i+1} is calculated. If $I_{i+1} \geq I_i$, the value of ε is modified and the calculation is repeated. When $I_{i+1} < I_i$, a controller has been determined which has reduced the index. Equation (6) is reevaluated and the iteration continues until the absolute value of $I_{i+1} - I_i$ is less then some η determined a priori or the total number of iterations exceeds some N.

b. Distributed Systems

For a system defined by equation (4) and index of performance (5), the algorithm for determining the change in controller is given by:

$$\delta u(x,t) = -\varepsilon \operatorname{sgn}\left[\int_\Omega d\tau \int_t^{t_f} d\tau\, W^T(\gamma,x,\tau,t) \frac{\partial F(\gamma,\tau)}{\partial v(\gamma,\tau)} + \frac{\partial F(x,t)}{\partial u(x,t)}\right] \tag{10}$$

The method of iteration is similar to case (a).

The derivation of equation (10) is presented in Appendix A. The derivation of equation (6) is similar and it is therefore omitted. Further discussion of the increment step, ε, is given in Appendix A.

EXAMPLES AND RESULTS

a. Lumped Parameter Problem - The algorithm (6) was first tested on a simple, first order, linear, limped parameter system for which a closed form solution of the optimal controller is available. The ordinary differential equation

$$\frac{dx}{dt}(t) + x(t) = u(t) \tag{11}$$

with its associated boundary condition

$$x(0) = 10 \tag{12}$$

has the integral representation:

$$x(t) = 10e^{-t} + \int_0^t e^{-(t-\tau)} u(\tau)\, d\tau \tag{13}$$

The index of performance is

$$I = 1/2 \int_0^{\sqrt{2}} (x(t)^2 + u(t)^2)\, dt \tag{14}$$

The associated algorithm for equations (13) and (14) is

$$\delta u(t) = - \varepsilon \operatorname{sgn}[\int_t^{\sqrt{2}} d\tau e^{-(\tau-t)} x(\tau) + u(t)] \tag{15}$$

A comparison of the exact optimal controller and the results of the algorithm is presented in Figure 2. The comparison is favorable.

b. Distributed System with Non-Distributed Input - The algorithm has been applied to a heat transfer problem, an insulated rod with the heat added at one end and cooled at the other. Heat is added uniformly at the surface x = 0, and the surface x = a is exposed to the ambient temperature, T_o. The system is described by the diffusion equation:

$$\frac{\partial v}{\partial t} (x,t) = \frac{k}{c\rho} \frac{\partial^2 v}{\partial x^2} (x,t); \; 0 < x < a \tag{16}$$

(where $v(x,t) = T(x,t) - T_o$ and $T(x,t)$ is the temperature).

with the boundary conditions

$$\frac{\partial v}{\partial x} (0,t) = - \frac{u(t)}{k} \tag{17}$$

and

$$\frac{\partial v}{\partial x} (a,t) = - \frac{h}{k} v(a,t) \tag{18}$$

The initial condition is zero. The integral equation representation for this system is

$$v(x,t) = \int_0^t W(x,t,\tau)\, u(\tau)\, d\tau \tag{19}$$

The function $W(x,t,\tau)$ is given in Appendix B.

It is desired to bring the surface at x = a to a temperature T_d. For this system the optimal controller (the rate of heat input, u(t)) must minimize the index of performance

$$I = \int_0^{50} (u(t) - u_{ss})^2 + (v_d - v(a,t))^2\, dt \tag{20}$$

where $v_d = T_d - T_o$ and u_{ss} is the heat rate at x = 0 necessary in steady state to maintain the desired temperature difference v_d. This index weighs both temperature error and excessive energy inputs. Henceforth in this example, all the variables are normalized with respect to v_d; thus we may write:

$$I = \int_o^{50} [(u(t) - 2.0)^2 + (1 - v(a,t))^2]\, dt \tag{21}$$

The results obtained on a digital computer are given in Fig. 3 for an aluminum rod (see Appendix B for the physical properties of materials mentioned). The iteration results are compared to approximations to analytic solutions (by truncation of infinite series) available in the literature [14]. The computations are repeated for carbon in Fig. 4.

For the aluminum rod example, the value of the index converges rapidly to the 4.766 area after 10 iterations. There is very little improvement in the index in the subsequent iterations. The following table, which includes the step size, demonstrates this.

ITERATION	INDEX	STEP SIZE
10	4.7666	1/256
11	4.7665	1/256
12	4.7664	1/512
13	4.7664	1/512
14	4.7663	1/1024
15	4.7663	1/1024
16	4.7663	1/2048
17	4.7663	1/2048
18	4.7663	1/4096

Up to this point we have required that the value of v(a,t) be as close as possible to the temperature difference 1 (really v_d). Let us assume that a response larger than 1 anywhere along the rod cannot be tolerated; thus, we introduce an additional restriction

$$v(x,t) \leq 1 \qquad \text{for } x \varepsilon\ [0,a] \tag{22}$$

This limitation is introduced into the problem through the index (equation (21)) by using the penalty function method.

The penalty function method used is that developed by Kelley [15] and Okamura [16]. For a given side constraint

$$v(x,t) \leq Vo$$

$$\text{or} \quad v(x,t) - Vo = g(x,t) \leq 0, \tag{23}$$

a penalty function may be formulated such that

$$p(g) \begin{cases} >0 \text{ for } g > 0 \\ =0 \text{ for } g \leq 0 \end{cases} \tag{24}$$

A non-negative penalty weighting coefficient, ϕ , is also introduced. The index is now modified to incorporate this penalty whenever (23) is not satisfied.

Equation (21) is rewritten

$$I = \int_0^a \int_0^{50} dxdt \left[\frac{1}{a} \quad (u(t)-2.0)^2 + \frac{1}{a} (1-v(a,t))^2 + \phi . p(x,t) \right], \tag{25}$$

where

$$p(x,t) = \begin{cases} 0 \text{ for } g \leq 0 \\ \left[\frac{v\ (x,t)}{Vo} \right]^{50} \text{ for } g > 0. \end{cases}$$

Thus, the third term of the integrand makes large contributions to the value of the index if relationship (23) is not satisfied. The index is now a double integral. The first two terms reduce to a single integral. The solution of the heat transfer optimization problem for a carbon rod with this additional restriction (equation (22)) is presented in Figs. 5 and 6. ϕ is 1.0 in this example.

c. Distributed System with a Distributed Input - The algorithm is applied to vibrations of a beam of finite length. Such a system (including a damping term) is described by the following partial differential equation.

$$\frac{\partial^4 y}{\partial \chi^4} + \frac{1}{a^2} \frac{\partial^2 y}{\partial t^2} + k \frac{\partial y}{\partial t} = \frac{P(x,t)}{EI} \tag{26}$$

The beam is freely hinged at its ends so that

$$y = \frac{\partial^2 y}{\partial \chi^2} = 0$$

$$\text{at } x = 0 \text{ and } x = L$$

With no initial displacement or velocity, the integral equation representation for the displacement of the system may be written

$$y(x,t) = \int_o^t \int_o^L W(x,\gamma,t,\tau\ ) \quad P(\gamma,\tau)\ d\,\gamma\ \ d\,\tau \tag{27}$$

where P is the loading per unit length (or the controller). Details of W are in Appendix C.

The controller must minimize the index of performance

$$I = \int_0^L \int_0^T \left[(P(x,t) - PP)^2 + (b_o(y(x,t) - yy))^2\right] dtdx \qquad (28)$$

where b_o is an arbitrary weighting constant. yy denotes the desired value of y(x,t) and PP is the corresponding value of control. yy and PP are known functions of x and t.

This example was tested using an iteration initialization of $P_o =$ 2600 lbs/ft. The iteration results at the midpoint of the beam are shown in Fig. 7. The value of the other constants and parameters used are listed in Appendix C.

SUMMARY AND DISCUSSION

The computer calculation time required per iteration is dependent upon the complexity of the example. In this work the values of the kernel functions were pre-calculated and stored in an array for all points needed in the integration routines. For the lumped parameter system, an average of .2 sec was required for each iteration. The operation time remains about .2 sec even for the distributed case with the nondistributed input provided we deal with the single integral index. But, when the penalty function is introduced and the index becomes a double integral, the time for a single iteration increases to about three-fourths of a minute. With the distributed input, single iteration times ran as great as three minutes. Of course, the only value in quoting these times lies in their relative magnitude. The times could be altered considerably by the speed of the computer and/or by more efficient programming techniques.

Algorithms that determine the optimal controllers for distributed systems described by integral equations have been presented. The algorithms are synthesized by a relaxation method based on first variations. The indices of performance may be single or multiple integrals. Additional constraints (in addition to the system equations) were introduced successfully into the problem using the penalty function technique. The results are considered satisfactory and compare well with those found in the literature.

For the examples considered, approximately twelve iterations of this first order optimal technique were necessary to generate a controller and response that remained essentially unchanged by further computation. Subsequent iterations produced changes of only several thousand per cent in the index. Convergence of first variations techniques similar to the one presented above is discussed by Merriam [2]. Okamura [16] has demonstrated the convergence of the penalty method applied to the optimal control problem for systems defined by differential equations. The proof for integral equations would be analogus.

It should be noted that the optimization technique given here could be extended to distributed systems described by nonlinear integral equations. The system equations in this text would have to be modified accordingly. Necessarily, much greater computational effort would be required for locating the optimal controller.

APPENDIX A

Derivation of Recursive Algorithm
For Distributed System
With Distributed Input

Expanding the integrand in the index of performance (equation (5)) into a Taylor's Series about the point u_o, v_o which satisfies equation (4) and neglecting the terms of order two and higher, we obtain

$$I \approx \int_\Omega dx \int_{t_o}^{t_f} dt \, [F(u_o(x,t), v_o(x,t)) + <\frac{\partial F(x,t)}{\partial v(x,t)}, \delta v(x,t)>$$

$$+ <\frac{\partial F(x,t)}{\partial u(x,t)}, \delta u(x,t)>] \tag{29}$$

where the derivatives $\frac{\partial F}{\partial v}$ and $\frac{\partial F}{\partial u}$ are evaluated at the point $u_o(x,t)$, $v_o(x,t)$. The notation $<\ldots,\ldots>$ referesto the inner product. Vectors δu and δv denote changes in u and v, i.e., $v = v_o + \delta v$ and $u = u_o + \delta u$. The pair δu and δv satisfy equation (4). ΔI, the approximate change in I is

$$\Delta I = \int_\Omega dx \int_{t_o}^{t_f} dt \, [<\frac{\partial F(x,t)}{\partial v(x,t)}, \delta v(x,t)> + <\frac{\partial F(x,t)}{\partial u(x,t)}, \delta u(x,t)>] \tag{30}$$

Substitute $\delta v(x,t)$ obtained from equation (4) into equation (30),

$$\Delta I = \int_\Omega dx \int_{t_o}^{t_f} dt \, [<\frac{\partial F(x,t)}{\partial v(x,t)}, \int_\Omega d\gamma \int_{t_o}^{t} d\tau \, W(x,\gamma,t,\tau)\delta u(\gamma,\tau)>$$

$$+ <\frac{\partial F(x,t)}{\partial u(x,t)}, \delta u(x,t)>] \tag{31}$$

Assuming that the order of integration may be interchanged in equation (31), we get

$$\Delta I = \int_\Omega dx \int_\Omega d\gamma \int_{t_o}^{t_f} dt \int_{t_o}^{t} d\tau <\frac{\partial F(x,t)}{\partial v(x,t)}, W(x,\gamma,t,\tau) \, \delta u(\gamma,\tau)>$$

$$+ \int_\Omega dx \int_{t_o}^{t_f} dt \, <\frac{\partial F(x,t)}{\partial u(x,t)}, \delta u(x,t)> \tag{32}$$

In the first term interchange the order of integration with respect to t and τ, and also x and γ, and then make a change of variables such that x and γ are interchanged along with t and τ. Equation (32) may now be expressed:

$$\Delta I = \int_\Omega dx \int_{t_o}^{t_f} dt \, <\int_\Omega d\gamma \int_{t}^{t_f} d\tau \, W^T(\gamma,x,\tau,t) \, \frac{\partial F(\gamma,\tau)}{\partial v(\gamma,\tau)} +$$

$$\frac{\partial F(x,t)}{\partial u(x,t)}, \delta u(x,t)> \tag{33}$$

Algorithm for the search of minimum I, and therefore the optimal controller, is obtained by requiring that δu be such that

$$\Delta I \leq 0 \tag{34}$$

ΔI can be made negative by letting:

$$\delta u(x,t) = -\varepsilon \operatorname{sgn} \left[\int_\Omega d\gamma \int_t^{t_f} d\tau \, W^T(\gamma,x,\tau,t) \frac{\partial F(\gamma,\tau)}{\partial v(\gamma,\tau)} + \frac{\partial F(x,t)}{\partial u(x,t)} \right] \tag{35}$$

where ε is some constant which regulates the magnitude of δu.

There is no unique determination of the step function ε. In order to test the acceptability of ε for any iteration, the values of the indices of performance between successive iterations are compared. ε is acceptable only when the inequality of expression (34) is satisfied. Convergence to an optimum controller is accomplished only if ε can be chosen so as equation (34) is fulfilled. The simplest function for selecting step size is a constant. A simple procedure would be to half ε each time (34) is not satisfied. But if it does satisfy (34) double the step for the next iteration. This technique is demonstrated in the flow chart (Figure 1). See Merriam [2] for a more complete and detailed discussion of convergence and step size.

APPENDIX B

Functions Used in Heat Transfer Example

The integral equation representation for this system is

$$v(x,t) = \int_0^t W(x,t,\tau)\ u(\tau)\ d\tau \tag{36}$$

where

$$W(x,t,\tau) = \frac{2\alpha}{k} \sum_{i=1}^{\infty} g(P_i,x)e^{-\alpha P_i^2(t-\tau)} \quad ,t>\tau \tag{37}$$

and

$$g(P_i,x) = \frac{P_i^2 + \beta^2}{a(P_i^2 + \beta^2) + \beta} \cos P_i x \tag{38}$$

The constants are:

$$\alpha = \frac{k}{c_\rho}$$

$$\beta = \frac{h}{k}$$

and P_i are ordered, positive roots of

$$P \tan (Pa) = \beta \tag{39}$$

The constant, h, and the length of the rod, a, are 2 BTU/HR-FT^2-$^\circ$F and .5 ft respectively.

Constants for different rod materials

(a) Aluminum

k = 117.0 BTU/HR-FT-$^\circ$F
c = .215 BTU/LB-$^\circ$F
ρ = .0915 LB/IN^3
α = 3.2297 FT^2/HR
β = .017094 1/FT

i	P_i
1	.181
2	6.2891
3	12.5683
⋮	
i	$\approx 2(i-1)\pi$, $i \geq 4$

(b) Carbon

K = 2.0 BTU/HR-FT-$^\circ$F
c = .18 BTU/LB-$^\circ$F
ρ = .05415 LB/IN^3
α = .1781 FT^2/HR
β = .6667 1/FT

i	P_i
1	1.09432
2	6.48798
3	12.6715
4	18.9200
5	25.1857
6	31.4583
7	37.7344
⋮	
i	$\approx 2\,(i-2)\,\pi,\ i \geq 8$

APPENDIX C

Functions Used In The Vibrations Example

The integral equation representation for this system is

$$y(x,t) = \int_0^t \int_0^L W(x,\gamma,t,\tau)\, P(\gamma,\tau)\, d\gamma\, d\tau \tag{40}$$

where $W(x,\gamma,t,\tau)$ is

$$W = \frac{2a^2}{LEI} \sum_{n=1}^{\infty} \frac{e^{-ka^2(t-\tau)/2}}{h(n)} \sin\left(\frac{n\pi x}{L}\right) \sin\left(\frac{n\pi\gamma}{L}\right) \sin(h(n)(t-\tau))$$

and $$h(n) = \left[\frac{a^2 n^4 \pi^4}{L^4} - \frac{a^4 k^2}{4}\right]^{1/2}$$

The constants are

E = modulus of elasticity
I = moment of inertia
L = length of beam
ω = weight of beam/unit length
k = damping coefficient
$a^2 = gEI/\omega$
g = gravitational constant

The iteration results shown in Figure 7 were executed using the following constants:

$EI = 10^3$ LB-FT^2
L = 1 FOOT
ω = .85 LB/FT
k = .002 SEC/FT^4
T = .0126 SEC
PP = 2000 LBS/FT
yy = .026 FT
b_o = 50000

ACKNOWLEDGEMENTS

The work described in this paper was partially supported by the National Science Foundation under grant number GK 2569. The valuable aid and encouragement of Melvin W. Crawford, Task Team Leader, of the U. S. Naval Ordnance Laboratory (White Oak, Maryland) are also gratefully acknowledged.

REFERENCES

1. M. Athans and P. Falb, Optimal Control, McGraw-Hill Book Co., New York, N. Y., 1966.

2. C. W. Merriam, Optimization Theory and the Design of Feedback Control Systems, McGraw-Hill Book Co., New York, N. Y., 1964.

3. A. G. Butkovskii and A. Y. Lerner, "The Optimal Control of Systems with Distributed Parameters", Automation and Remote Control, Vol. 21, No. 6, June 1960, pp. 682-691.

4. A. G. Butkovskii, "Optimum Processes in Systems with Distributed Parameters", Automation and Remote Control, Vol. 22, No. 1, Jan. 1961, pp. 17-26.

5. A. G. Butkovskii, "The Maximum Principle for Optimum Systems with Distributed Parameters", Automation and Remote Control, Vol. II, No. 10, Oct. 1961, pp. 1288-1301.

6. A. G. Butkovskii, "Some Approximate Methods for Solving Problems of Optimal Control of Distributed Parameter Systems", Automation and Remote Control, Vol. 22, No. 12, Dec. 1961, pp. 1565-1575.

7. A. G. Butkovskii, "The Broadened Principle of the Maximum for Optimal Control Problems", Automation and Remote Control, Vol. 24, No. 3, Mar. 1963, pp. 314-327.

8. K. A. Lurie, "On the Hamilton-Jacobi Method in Variational Problems of Partial Differential Equations", Journal of Applied Mathematics and Mechanics, Vol. 27, No. 2, 1963, pp. 255-264.

9. K. A. Lurie, "The Mayer-Bolza Problem for Multiple Integrals and the Optimization of the Performance of Systems with Distributed Parameters", Journal of Applied Mathematics and Mechanics, Vol. 27, No. 5, 1963, pp. 842-853.

10. R. Bellman, Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957.

11. R. Bellman and S. E. Dreyfus, Applied Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1962.

12. J. L. Lions, Controle Optinal de Systemes Gouvernes par des Equations Aux Derivees Partielles, Dunod-Gauthier-Villass, Paris, 1968. (English translation by S. K. Mitter, published by Springer-Verlag, 1969).

13. Control of Distributed Parameter Systems, A Symposium of the American Automatic Control Council, JACC, Boulder, Colorado, August 6, 1969 (published by ASME).

14. H. C. Khatri and R. E. Goodson, "Optimal Feedback Solutions for a Class of Distributed Systems", Transactions of the ASME, Journal of Basic Engineering, Vol. 88, June 1966, pp. 337-342.

15. H. J. Kelley, "Method of Gradients", Optimization Techniques with Applications to Aerospace Systems, G. Leitmann (Editor), Academic Press, N. Y., N. Y., 1962, Chap. 6, pp. 205-254.

16. K. Okamura, "Some Mathematical Theory of the Penalty Method for Solving Optimum Control Problems", J. SIAM Control, Ser. A, Vol. 2, No. 3, 1964, pp. 317-331.

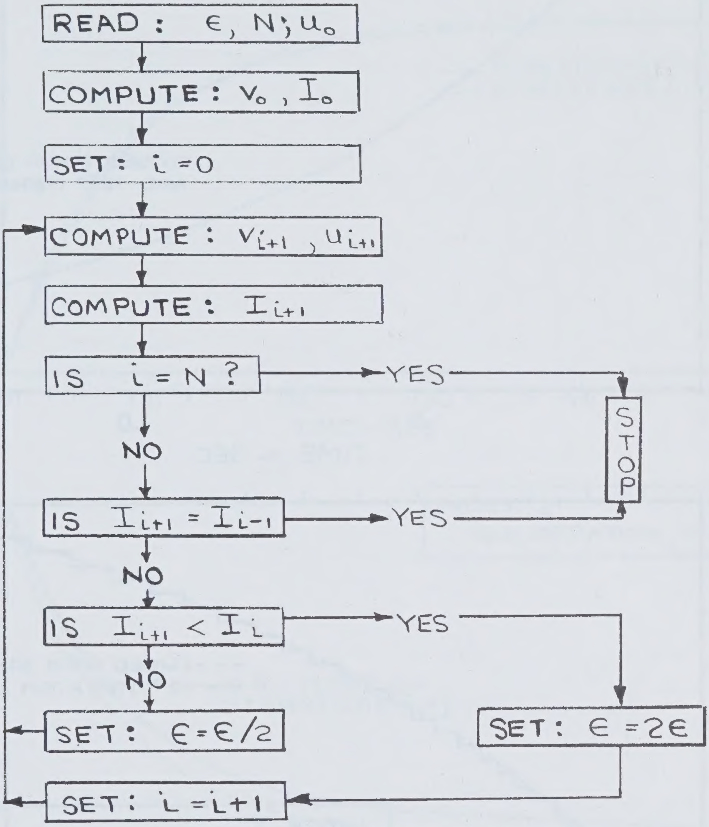

READ : ∈, N; u_0
COMPUTE : V_0 , I_0
SET : i = 0
COMPUTE : V_{i+1} , u_{i+1}
COMPUTE : I_{i+1}
IS i = N ?
YES
STOP
NO
IS $I_{i+1} = I_{i-1}$
YES
NO
IS $I_{i+1} < I_i$
YES
NO
SET : ∈ = ∈/2
SET : ∈ = 2∈
SET : i = i+1

FIGURE 1: FLOW CHART

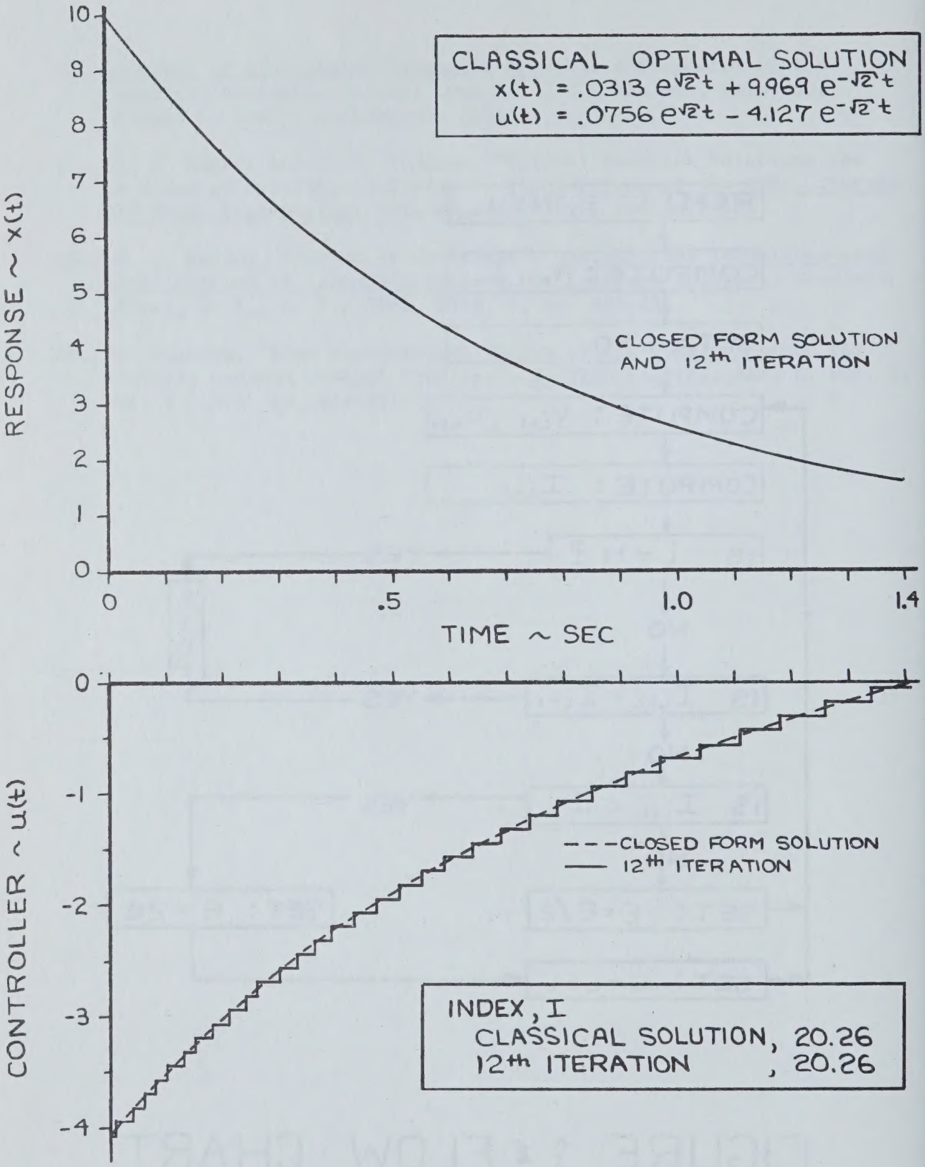


FIGURE 2

OPTMAL CONTROLER AND RESPONSE
FOR A LUMPED PARAMETER SYSTEM

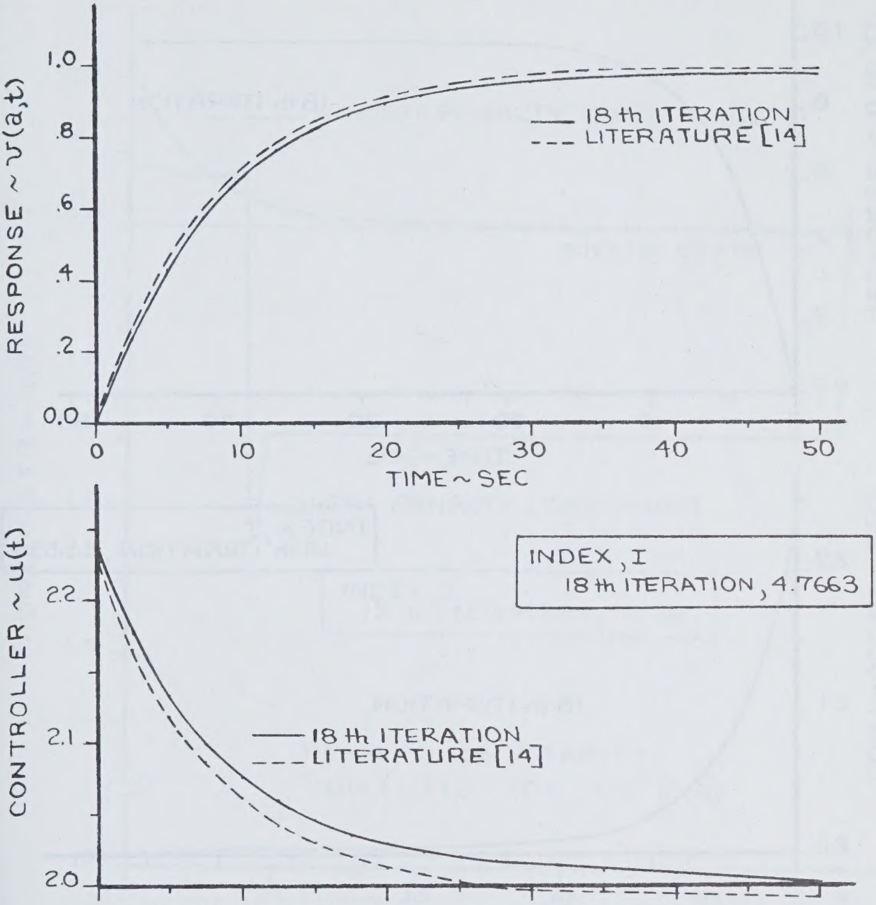


FIGURE 3
HEAT-TRANSFER EXAMPLE,
ALUMINUM ROD,
SINGLE INTEGRAL INDEX

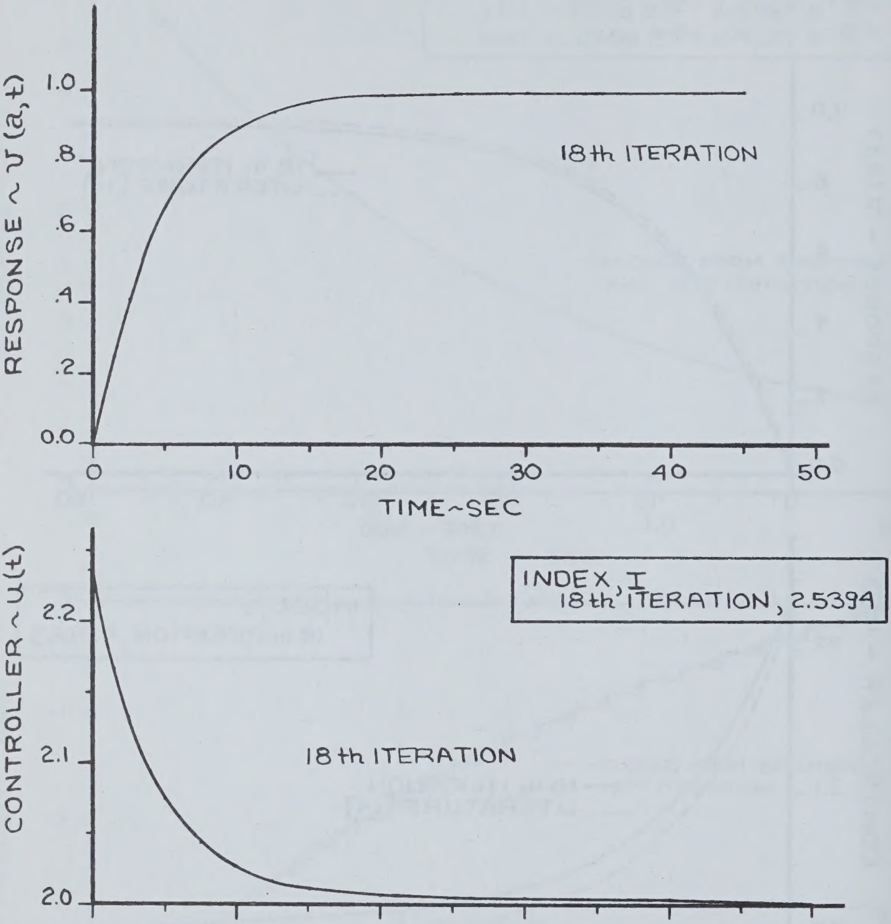


FIGURE 4

HEAT-TRANSFER EXAMPLE
CARBON ROD
SINGLE INTEGRAL INDEX

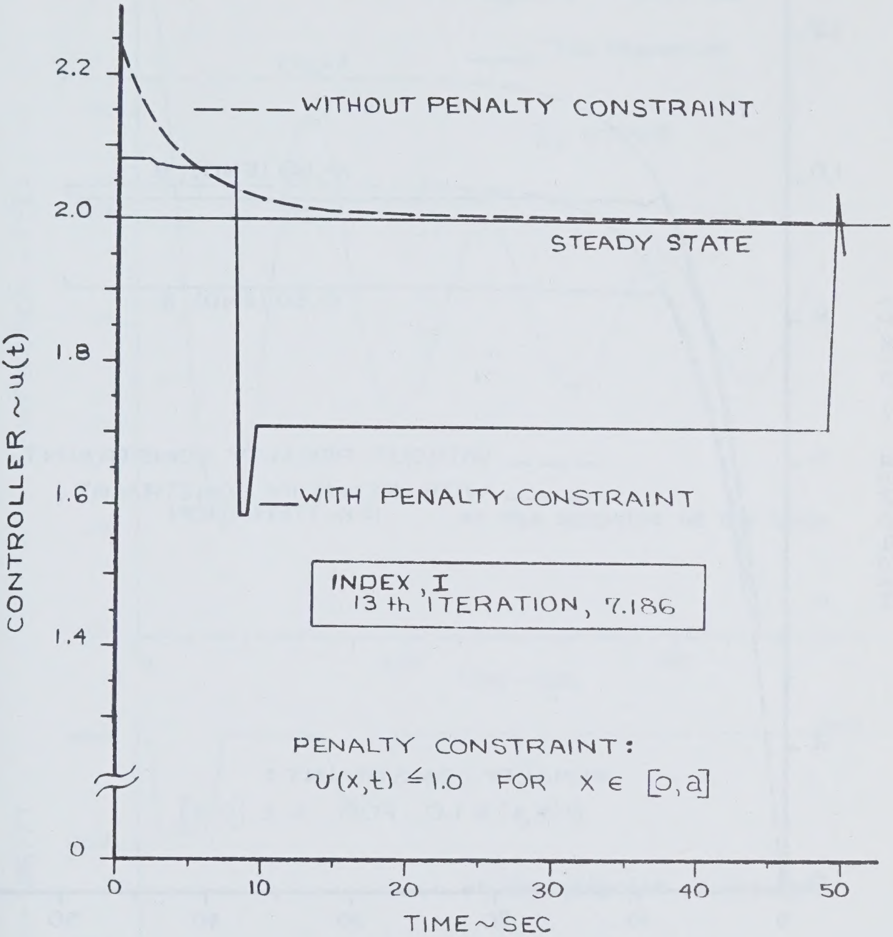


FIGURE 5

PENALTY-FUNCTION EXAMPLE, CONTROLLER FOR A CARBON ROD, MULTIPLE INTEGRAL INDEX.

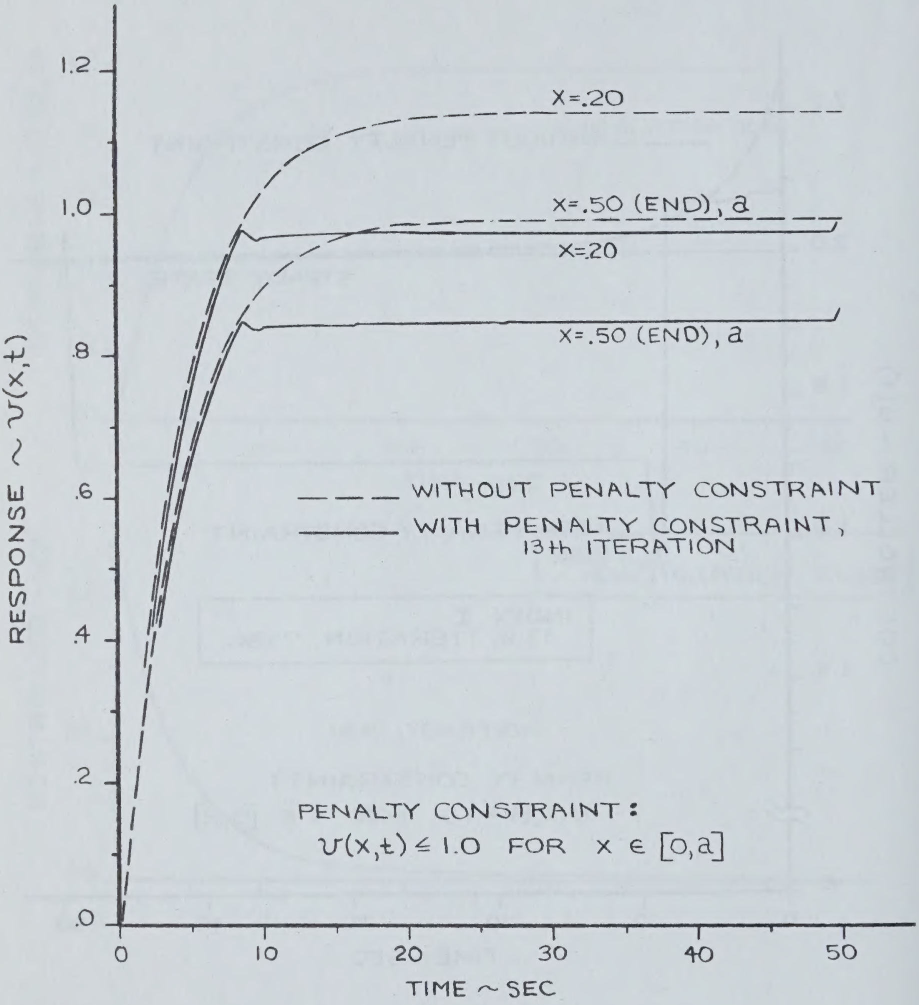


FIGURE 6
PENALTY-FUNCTION EXAMPLE, RESPONSE FOR A CARBON ROD, MULTIPLE INTEGRAL INDEX.

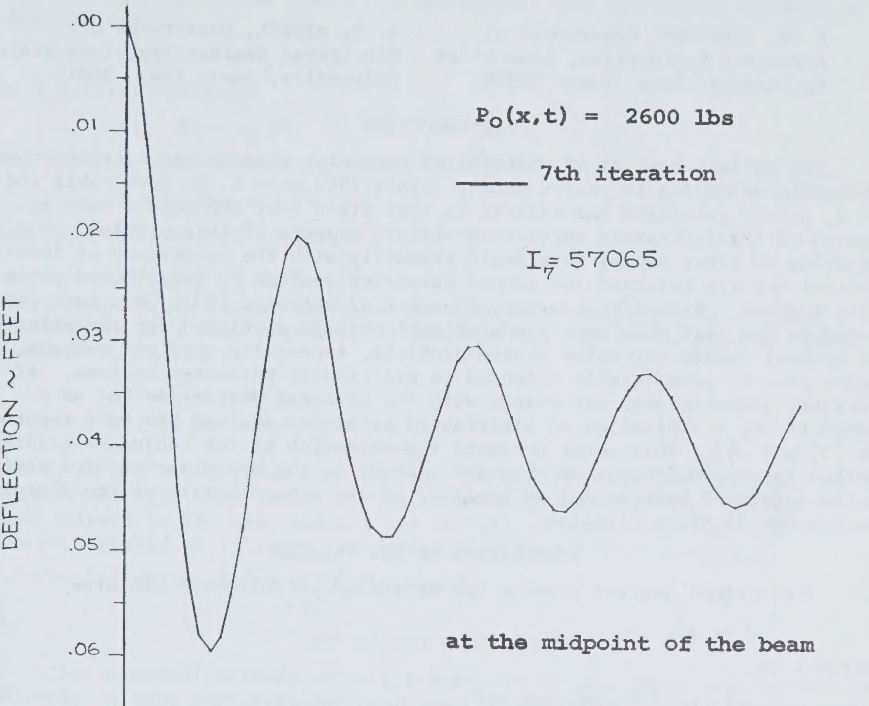


FIGURE 7
DISTRIBUTED SYSTEM WITH A DISTRIBUTED CONTROLLER

OPTIMIZATION OF PARABOLIC AND HYPERBOLIC DISTRIBUTED SYSTEMS VIA THE CONJUGATE GRADIENT METHOD

D. E. CORNICK*, Department of Aerospace Engineering, Iowa State University, Ames, Iowa 50010

A. N. MICHEL, Department of Electrical Engineering, Iowa State University, Ames, Iowa 50010

INTRODUCTION

The optimal control of distributed parameter systems has received considerable attention in recent years. Since 1960 when A. G. Butkovskii and A. V. Lerner initiated the efforts in this field over 200 papers have appeared in the literature concerning various aspects of this problem. The majority of these papers have dealt primarily with the extensions of theoretical results obtained for lumped parameter systems to distributed parameter systems. Recently, however, a number of authors, [1] – [4], have alluded to the fact that some computational methods developed for the solution of optimal lumped parameter system problems, especially gradient methods, might also be beneficially extended to distributed parameter systems. At present, computational experience with the steepest descent method as related to the optimization of distributed parameter systems has been reported in [3] and [4]. This paper presents the extension of the conjugate gradient method (a second-generation gradient method) to the optimization of distributed parameter systems and an analysis of the suboptimality of the approximation due to discretization.

FORMULATION OF THE PROBLEM

The optimal control problem may be stated as follows: minimize

$$J[u;\ x], \tag{1}$$

subject to

$$\psi[u;\ x] \geq \theta,\ x \in X \text{ and } u \in U, \tag{2}$$

where (with $\Omega \subset R^m$, $T \subset R^1$) $X \subset L^2[\Omega \times T]$ is called the state space, $U \subset L^2[\Omega \times T]$ the set of admissible controls, and $J[u;\ x]$ is a real-valued functional defined on the product space $U \times X$. The nonlinear operator ψ is defined on $U \times X$, and θ is the null vector of this product space. The functional $J[\cdot;\cdot]$ will be referred to as the cost index, and the conditions of (2) will be called the constraints. As a consequence of the constraints, the state trajectory $x(t)$ is dependent on the control u.

Let $J[u;\ x]$ be a real-valued quadratic functional defined on the real separable Hilbert space $U \times X$, generated by the self-adjoint operators M and N, and by the inner product $<\cdot;\cdot>$, and let $J[u;\ x]$ be specified by

$$J[u;\ x] = c_0 + <c_1,\ u> + \frac{1}{2} <u,\ Mu> + <c_2,\ x> + \frac{1}{2} <x,\ Nx> \tag{3}$$

or

$$J\left[\begin{bmatrix} u_d \\ u_b \end{bmatrix};\ x\right] = c_0 + <\begin{bmatrix} c_1^d \\ c_1^b \end{bmatrix}, \begin{bmatrix} u_d \\ u_b \end{bmatrix}> + \frac{1}{2} <\begin{bmatrix} u_d \\ u_b \end{bmatrix}, [M_d | M_b]\begin{bmatrix} u_d \\ u_b \end{bmatrix}>$$
$$+ <c_2,\ x> + \frac{1}{2} <x,\ Nx>, \tag{4}$$

where $c_0 \in R^1$, $c_1 \in U$, $c_2 \in X$, and where the vector c_1 and the operator M are partitioned into $\begin{bmatrix} c_1^d \\ c_1^b \end{bmatrix}$ and $[M_d | M_b]$, respectively. At any time $t \in T$ the distributed state of the system is denoted by $x(r,\ t)$, the distributed control vector is denoted by $u_d(r,\ t)$, and the boundary control vector is represented by $u_b(r_b,\ t)$. The ith component of $x(r,\ t)$ is written as $x_i(r_1,\ r_2,\ \cdots,\ r_m,\ t)$, $i = 1,\ 2,\ \cdots\cdots,\ n$; the ith component of $u_d(r,\ t)$ is denoted by $u_d^i(r_1,\ r_2,\ \cdots,\ r_m,\ t)$, $i = 1,\ 2,\ \cdots,\ p \leq n$; and the ith component of

$u_b(r_b, t)$ is represented as $u_b^i(r_b^1, r_b^2, \cdots, r_b^m, t)$. Here $r_b \epsilon \partial\Omega$, the boundary of Ω.

Let the constraint $\psi[u; x]$ be decomposed into the linear distributed dynamical system

$$Sx(r, t) = u_d(r, t) \tag{5}$$

with initial condition

$$x(r, 0) = x_0(r) \tag{6}$$

and terminal condition (target set)

$$\Psi_\ell x(r, T_f) = x_D(r), \tag{7}$$

and into the boundary condition

$$Tx(r_b, t) = u_b(r_b, t), \tag{8}$$

where S and T are linear differential operators consisting of a linear combination of a time differential operator D_t, and a spatial differential operator D_r, given by

$$S \equiv c_1 D_t + c_2 D_r , \tag{9}$$

$$T \equiv [c_3 D_t + c_4 D_r]\Big|_{\partial\Omega} , \quad c_i \epsilon R^1 , \tag{10}$$

and Ψ_ℓ is a n × n self-adjoint matrix, and $x_D(r)$ is the desired state at the final time. The distributed optimal control problem for (5) through (8) with respect to the cost index J and the set of admissible controls U can now be restated as follows: determine the control $u^* \epsilon U$ such that

$$J[u^*; x(u^*)] = \min_{u \epsilon U} \left\{ J[u; x(u)] \right\}. \tag{11}$$

OPTIMALITY CONDITIONS

The numerical methods which are employed in this paper are directly applicable to only unconstrained problems. The penalty function method will be utilized to render the constrained problem amenable to gradient methods. The original problem is then replaced by an approximate unconstrained problem. In the subsequent development the following penalty function will be used:

$$P[x(r, T_f)] = \langle \Psi, W\Psi \rangle = \langle \Psi_\ell x(r, T_f) - x_D(r), W[\Psi_\ell x(r, T_f) - x_D(r)] \rangle, \tag{12}$$

where W is an n × n self-adjoint matrix of penalty constants.

The constrained problem may be restated as an unconstrained problem as follows: minimize

$$J_p[u; x] = J[u; x] + P[x(r, T_F)] \tag{13}$$

subject to (5), (6), and (8). In the unconstrained problem the dynamical system is not a constraint, but rather a side condition, which must be solved only to evaluate the penalized cost index J_p.

By representing the solution of the state system in terms of appropriate Green's functions it is possible to remove the explicit dependence of the cost index on the state trajectory. Thus, let the solution of the dynamical system exist, be unique, and have the representation

$$x(r, t) = \Phi(t)x_0(r) + S^{-1}(t)u_d(r, t) + T^{-1}(t)u_b(r_b, t), \tag{14}$$

where $\Phi(t)x_0(r)$ denotes contribution to the solution at time t due to the initial conditions, $S^{-1}(t)u_d(r, t)$ denotes contribution to the solution at time t due to the distributed control, and $T^{-1}(t)u_b(r_b, t)$ denotes contribution to the solution at time t to the boundary control. The state of the system at time T_f is evaluated as

$$x(r, T_f) = \Phi(T_f)x_0(r) + S^{-1}(T_f)u_d(r, t) + T^{-1}(T_f)u_b(r, t). \tag{15}$$

The state trajectory is eliminated from the penalized cost index by substituting (14) and (15) into (13), i.e.,

$$\begin{aligned} J_p[u; x] &= J[u;\ \Phi(t)x_0 + S^{-1}(t)u_d + T^{-1}(t)u_b] + \\ &\quad + P[\Phi(T_f)x_0 + S^{-1}(T_f)u_d + T^{-1}(T_f)u_b]. \end{aligned} \tag{16}$$

Simplification of (16) yields the standard quadratic form

$$J_p[u] = J_0 + <c, u> + \frac{1}{2} <u, Au>, \tag{17}$$

where

$$\begin{aligned} J_0 = c_0 &+ <\left\{\Phi^*(t)c_2 - \Phi^*(T_f)\Psi_\ell^*[W^* + W]\right\}x_D,\ x_0> + <[\tfrac{1}{2}\ \Phi(t) \\ &+ \Psi_\ell\Phi(T_f)]x_0, [\tfrac{1}{2}\ N\Phi(t) + W\Psi_\ell\Phi(T_f)]x_0> + <x_D, Wx_D>, \end{aligned} \tag{18}$$

$$c = \begin{bmatrix} c_1 \\ \\ c_2 \end{bmatrix} = \begin{bmatrix} c_1^d + [S^{-1}(t)]^*c_2 + \left\{\frac{1}{2}[S^{-1}(t)]^*(N^* + N)\right\}\Phi(t)x_0 + \\ \left\{[S^{-1}(T_f)]^*\Psi_\ell^*[W^* + W]\Psi_\ell\right\}\Phi(T_f)x_0 - [S^{-1}(T_f)]^*\Psi_\ell^*[W^* + W]x_D \\ \\ c_1^b + [T^{-1}(t)]^*c_2 + \left\{\frac{1}{2}[T^{-1}(t)]^*(N^* + N)\right\}\Phi(t)x_0 + \\ \left\{[T^{-1}(T_f)]^*\Psi_\ell^*[W^* + W]\Psi_\ell\right\}\Phi(T_f)x_0 - [T^{-1}(T_f)]^*\Psi_\ell^*[W^* + W]x_D \end{bmatrix}, \tag{19}$$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ \\ A_{21} & A_{22} \end{bmatrix}, \tag{20}$$

with

$$A_{11} = M_d + [S^{-1}(t)]^*NS^{-1}(t) + 2[S^{-1}(T_f)]^*\Psi_\ell^*W\Psi_\ell S^{-1}(T_f), \tag{21}$$

$$A_{12} = [S^{-1}(t)]^*NT^{-1}(t) + 2[S^{-1}(T_f)]^*\Psi_\ell^*W\Psi_\ell T^{-1}(T_f), \tag{22}$$

$$A_{21} = [T^{-1}(t)]^*NS^{-1}(t) + 2[T^{-1}(T_f)]^*\Psi_\ell^*W\Psi_\ell S^{-1}(T_f), \tag{23}$$

$$A_{22} = M_b + [T^{-1}(t)]^*NT^{-1}(t) + 2[T^{-1}(T_f)]^*\Psi_\ell^*W\Psi_\ell T^{-1}(T_f). \tag{24}$$

The necessary condition for u^* to be the element of U which minimizes J_p is that the gradient (utilizing the Fréchet derivative) of J_p with respect to the control u vanish at u^*. Thus

$$\left.\frac{\partial J_p}{\partial u}\right|_{u^*} = g(u^*) = \left.\begin{bmatrix} \dfrac{\partial J_p}{\partial u_d} \\ \dfrac{\partial J_p}{\partial u_b} \end{bmatrix}\right|_{u^*} = c + \frac{1}{2}[A + A^*]u^* = 0. \tag{25}$$

If A is a self-adjoint operator, then (25) yields

$$g(u^*) = c + Au^* = 0. \tag{26}$$

From (21) through (24) it follows that for A to be self-adjoint, M_d, M_b, N, Ψ_ℓ, and W must be self-adjoint. If A is self-adjoint then the Hessian of J_p given by

$$\frac{\partial^2 J_p}{\partial u^2} = A, \tag{27}$$

is positive definite. Consequently, (26) and (27) are necessary and sufficient conditions for u^* to exist and be the unique optimal control for the penalized cost index J_p.

CONJUGATE GRADIENT METHOD

The conjugate gradient method is an iterative approach of minimizing the cost functional J(u). It is based on the computation of the gradient

$g(u)$ of $J(u)$. The conjugate gradient method seeks to find the minimizing element u^* by generating a sequence $\{u_n\}$ which satisfies at each step the problem constraints and the optimality conditions, and iterates on $g(u)$ such that $\{u_n\} \to u^*$. It has been shown in [5] that if $J[u]$ is a quadratic functional defined on a real Hilbert space, then the conjugate gradient method does indeed generate such a sequence. This result includes as a special case the class of distributed parameter systems considered presently.

The conjugate gradient algorithm may be summarized as follows: (1) for $n = 0$, guess the initial control function u_0; (2) integrate (by some appropriate finite difference method) the state system from t_0 to t_f; (3) integrate the co-state system backwards from t_f to t_0; (4) calculate the gradient $g(u_n) = g_n$; (5) calculate the conjugate gradient parameter β_n, where

$$\beta_n = \langle g_n, g_n\rangle / \langle g_{n-1}, g_{n-1}\rangle, \; \beta_0 = 0 \text{ if } n = 0; \tag{28}$$

(6) calculate the direction of search s_n, where

$$s_n = - g_n + \beta_n s_{n-1} ; \tag{29}$$

(7) iterate on the control

$$u_{n+1} = u_n + \gamma_n s_n \triangleq C_G(u_n), \tag{30}$$

where γ_n is determined by performing a one-dimensional minimization along the direction of search; and (8) increase n and repeat steps 2 through 7 until specified convergence criteria are satisfied.

DISCRETIZATION AND SUBOPTIMALITY OF THE APPROXIMATE SOLUTION

The conjugate gradient method is an iterative procedure and is therefore only practical when programmed on a high-speed digital computer. Thus the original continuous problem is actually replaced by a discrete problem. In the process of transforming the continuous problem into its discrete analog, a number of approximations are made which include errors. Basically, two types of approximations are involved: (1) approximations to elements of a Hilbert space (e.g., the approximation of continuous functions by piecewise continuous polynomials); (2) approximations of operators defined on a Hilbert space (e.g., approximations of differential and integral operators by finite-difference and summation operators, respectively). In addition, there are always errors encountered which are due to numerical roundoff. In the present case it was found that roundoff errors are negligible when compared to other approximations involved (double precision programming was utilized). Furthermore, it became evident that the approximations of the integral operators (e.g., inner products, etc.) could be made several orders of magnitude more accurate than the approximation of the differential operators. Thus, errors due to quadrature formulas were also found to be negligible. Hence, the two primary error sources were found to be: (1) approximations of functions; (2) approximations of differential operators.

If $u_{n+1} = C_G(u_n)$ represents the exact conjugate gradient iteration, then the approximations discussed above yield what will be referred to as the approximate conjugate gradient method $\tilde{u}_{n+1} = \tilde{C}_G(\tilde{u}_n)$. The following important questions arise: (1) given that the convergence of the exact conjugate gradient method is assured, under what conditions will there result convergence of the approximate conjugate gradient method? (2) if the approximate iteration does converge numerically (in general $\{\tilde{u}_n\} \to \tilde{u}^* \neq u^*$), at what step should the iteration be terminated in order to insure a reasonable estimate to u^*? (3) how is this estimate to u^* made, and how suboptimal is $\tilde{u}^*$? Before answering these questions some additional nomenclature and definitions have to be introduced. Let $\tilde{u}_n \in \tilde{U}$ denote the numerical approximation to $u \in U$, where $\tilde{U}$ is a subspace of U, let $\tilde{J}$ be the numerical approximation to J, and let $\|\cdot\|$ represent the norm of a vector.

Definition 1. [6] If there exists a set $S = \left\{ \tilde{u}: \ \tilde{u}_{n+1} = \tilde{u}_n, \ n \geq N \right\}$, then $\left\langle \tilde{u}_n \right\rangle$ is said to be numerically convergent and S is said to be the state of numerical convergence.

Definition 2. Let $J[u^*] = \min_{u \in U} J[u]$, then the optimal control error $\| e_u \|$ is defined by $\| e_u \| = \| \tilde{u}^* - u^* \|$, where $\tilde{u}^* = \tilde{u}_N \in S$.

Definition 3. The cost functional error e_J is defined as $e_J = | J[u^*] - \tilde{J}[\tilde{u}^*] |$.

Since $J[u^*]$ cannot be computed, the cost functional error e_J is a measure of the suboptimality of the approximate solution.

It is shown in [7] that from a computer storage and from computer time considerations, only moderately accurate finite difference methods for the solution of the state and co-state systems are practical. Since errors introduced by the finite-difference solution of the state and co-state systems cause errors in computing the gradient vector, it becomes necessary to consider the effects of gradient errors on the approximate conjugate gradient operator.

Gradient errors have two effects on the operator $\tilde{C}_G$: (1) direction of search errors; (2) linear minimization errors. Until more accurate finite difference methods are developed, it appears that direction of search errors must be tolerated. However, linear minimization errors can be avoided when the effects of gradient errors on this phase of the iteration are understood.

Let the exact gradient of the cost functional $J[\cdot]$ at $\tilde{u}_n$ be given by $g(\tilde{u}_n)$, then

$$g(\tilde{u}_n) = \tilde{g}(\tilde{u}_n) + e_g(\tilde{u}_n; h), \tag{31}$$

where $\tilde{g}(\tilde{u}_n)$ is the approximate gradient at $\tilde{u}_n$, and $e_g(\tilde{u}_n; h)$ is the gradient error which, as indicated, depends on the net size parameter, h, of the finite difference method used to compute solutions of the state and co-state systems. The proofs of the results which follow are given in [7].

Lemma 1. If the dynamical system is linear and if J is quadratic, then the truncation error in the gradient, $e_g(\tilde{u}_n; h)$ is linear in $\tilde{u}_n$. Specifically, $e_g(\tilde{u}_n; h) = E_h \tilde{u}_n + c_g$, where E_h is a linear operator depending on the net size parameter h, and $c_g \in \tilde{U}$.

Theorem 1. If the dynamical system is linear, and if J is a quadratic functional, then $\tilde{g}(\tilde{u}_n)$ is the exact gradient of the quadratic functional

$$\hat{J} = J_0 + \langle c - c_g, u \rangle + \frac{1}{2} \langle \tilde{u}, (A - E_h)\tilde{u} \rangle.$$

Let $\tilde{s}(\tilde{u}_n) = \tilde{s}_n$ be the approximation to s_n defined in (29). Then it is easily shown that

$$\tilde{\gamma}_n = - \langle g_n, \tilde{s}_n \rangle / \langle \tilde{s}_n, A\tilde{s}_n \rangle \tag{32}$$

and

$$\hat{\gamma}_n = - \langle \tilde{g}_n, \tilde{s}_n \rangle / \langle \tilde{s}_n, \tilde{A}\tilde{s}_n \rangle, \ \tilde{A} = A - E_h \tag{33}$$

denote the step size parameters defined by (30) in the direction $\tilde{s}_n$ for J and $\hat{J}$, respectively. For the problems considered, the operators A and $\tilde{A}$ are, respectively, double integral and double summation operators. Consequently, (32) and (33) are not generally used in practice. Instead, the following methods are usually utilized to determine γ_n: (i) cubic interpolation based on functional and directional derivative information [8]; (ii) quadratic interpolation based on functional information only; (iii) linear gradient interpolation based on directional derivative information only. In [7] it is shown that with gradient errors, (a) method (i) does not converge to $\tilde{\gamma}_n$ nor to $\hat{\gamma}_n$; (b) method (ii) converges to $\tilde{\gamma}_n$; (c) method (iii) converges to $\hat{\gamma}_n$. With these results, the following theorem can be proved [7].

Theorem 2. If the conditions of Theorem 1 are satisfied, and if A is a self-adjoint operator, and if method (iii) above is used in the linear minimization, then the iteration $\tilde{u}_{n+1} = \tilde{C}_G(\tilde{u}_n)$ generates a sequence $\{\tilde{u}_n\}$ which converges to the minimizing element of $\hat{J}$. Specifically, $\hat{J}[\tilde{u}_N] = \min_{\tilde{u}\in\tilde{U}} \hat{J}[\tilde{u}]$.

Remark. Theorem 2 implies that the sequence $\{\tilde{u}_n\}$ generated by $\tilde{C}_G$ minimizes $\hat{J}$ and not J nor $\tilde{J}$. Since $\{\tilde{u}_n\}$ does not minimize J, it is desirable to compute the optimal control error; however, this is impossible without a prior knowledge of the solution u^*. Nevertheless, estimates of the optimal control error may be obtained by the results given below.

Lemma 2. Let $\nabla_u J$ denote the gradient of a positive definite quadratic functional J defined on a real Hilbert space H. Then (i) $J_g[u] = \|\nabla_u J\|^2 = \|g(u)\|^2$ is also a quadratic functional with Hessian 2AA, (ii) the set defined by $S_g = \{u: \|g\|^2 = c\}$ is a hyperellipsoid, and (iii) if $J_g[u_g^*] = \min_{u\in U} J_g[u]$, then $u_g^* = u^*$.

Lemma 3. Consider the translation defined by $\hat{u} = u - u^*$. Then (i) $\hat{J}[\hat{u}] = J[u^*] + \frac{1}{2} <\hat{u}, A\hat{u}>$, and (ii) $\nabla_{\hat{u}} \hat{J}[\hat{u}] = \nabla_u J[u]$.

Theorem 3. The vectors defined by $\|\hat{u}_u\|^2 = \max_{\hat{u}\in S_{\hat{g}}} \|\hat{u}\|^2$ and $\|u_1\|^2 = \min_{\hat{u}\in S_{\hat{g}}} \|\hat{u}\|^2$ are eigenvectors of A^2 with eigenvalues M_A^2 and m_A^2, respectively.

An estimate of the optimal control error can now be given in terms of the gradient vector norm.

Theorem 4. Let $\|g(\tilde{u}_N)\|^2$ denote the exact normed gradient squared of J at $\tilde{u}_N$. Then

$$\|g(\tilde{u}_N)\|^2/M_A^2 < \|\tilde{u}_N - u^*\|^2 < \|g(u_N)\|^2/m_A^2 .$$

From Theorem 2, $\tilde{g}[\tilde{u}_n] \to 0$ with n. This results in a method by means of which $\|g(\tilde{u}_N)\|^2$ can be estimated.

Theorem 5. In the state of numerical convergence (see Definition 1), the approximate conjugate method $\tilde{u}_{n+1} = \tilde{C}_G(\tilde{u}_n)$, with $\hat{\gamma}_n = <\tilde{g}_n, \tilde{s}_n>/<\tilde{s}_n, \tilde{A}s_n>$, insures that $\tilde{g}(\tilde{u}_N) = 0$. Thus $\|g(\tilde{u}_N)\|^2 = \|e_g(u_N; h)\|^2$.

By combining Theorems 4 and 5 an estimate of the optimal control error is obtained by the following Lemma.

Lemma 4. Let $e_g(\tilde{u}_N; h)$ be the gradient error at the Nth iteration of $\tilde{u}_{n+1} = \tilde{C}_g(\tilde{u}_n)$. Then the estimate

$$\|e_g(\tilde{u}_N; h)\|/M_A < \|\tilde{u}^* - u^*\| < \|e_g(\tilde{u}_N; h)\|/m_A$$

holds.

An estimate of the cost functional error can also be given in terms of the gradient error [7].

Theorem 6. Let J be a quadratic functional defined **on U** generated by a self-adjoint operator A and by an inner product $<\cdot;\cdot>$. Let $\tilde{J}$ be the discrete approximation of J defined on the subspace $\tilde{U}$ **generated** by $\tilde{A}$ and by the inner product $(\cdot;\cdot)$. Then

$$|J[u^*] - \tilde{J}[\tilde{u}^*]| \leq \frac{1}{2} \sum_1^2 |e_{sp}^i| + \frac{1}{2} \|\tilde{u}^*\| [\|e_c\| + \|e_g\|] + |e_{J_0}| + \frac{1}{2} \{\|\tilde{c}\| + \|e_c\| + \|e_g\|\} \frac{\|e_g\|}{m_A} ,$$

where $J_0 = \tilde{J}_0 + e_{J_0}$, $c = \tilde{c} + e_c$, and $<\cdot;\cdot> = (\cdot;\cdot) + e_{sp}^i$.

Since both estimates of the optimal control error and the cost functional error are based on the gradient error, a method for estimating the gradient error norm is required. At present two methods have been utilized for estimating $\|e_g\|$: (1) error bounds in terms of higher order differences; (2) asymptotic extrapolation. Since the first method is rather conservative, only the second method is discussed here. Asymptotic extrapolation is an attempt to actually compute the gradient error, and is based on the fact that if the approximation to the gradient is of order p, then

$$g(\tilde{u}_N) = \tilde{g}(u_N; h) + h^p \tilde{e}_g(\tilde{u}_N) + o(\tilde{u}_N; h^{p+1}),$$

where $\tilde{e}_g$ is defined as the magnified error function. Solving for the gradient at $\tilde{u}_N$ by using step sizes of h and qh, respectively, $(0 < q < 1)$, two equations in $g(\tilde{u}_N)$ and $\tilde{e}(\tilde{u}_N)$ result which, when solved, yield

$$||e_g(\tilde{u}_N;\, h)|| \cong [1/(1 - q^p)]||\tilde{g}(\tilde{u}_N;\, qh) - \tilde{g}(\tilde{u}_N;\, h)||.$$

In view of Theorem 2, $\tilde{g}(\tilde{u}_N;\, h) = 0$, and thus the estimate

$$||e_g(\tilde{u}_N;\, h)|| \cong [1/(1 - q^p)]||\tilde{g}(\tilde{u}_N;\, qh)||$$

for the gradient error norm is obtained.

The error estimates given above require much less computer time than the standard method for this type of error estimates (which is to resolve the entire problem with a refined mesh and then extrapolate each of the parameters for which the error estimates are desired).

NUMERICAL RESULTS

All computations reported were performed on the IBM 360 Model 65 digital computer using the FORTRAN IV language and the WAT IV compiler with double-precision arithmetic. Computation times quoted are the time used by the Central Processing Unit during program execution. Although the Central Processing Unit time is the best measure of the computing effort required, it is not precisely reproducible on identical programs due to the multi-programming feature of the system. Storage requirements reported are in terms of array area used in BYTES, which does not include object code storage requirements.

The solution of the state and co-state systems were performed with a standard second order symmetric finite-difference algorithm [9]. The multiple quadrature algorithm used to compute the cost functional and the inner products was based on the Gauss-Legendre 10-point quadrature formulas [10]. Piecewise continuous quadratic polynomials were used to obtain function values between the net points. Each linear minimization required in the solution was based on a linear interpolation of the approximate directional derivative along the direction of search.

Example 1. Distributed Control of the One-Dimensional Heat Equation

The unconstrained, fixed time, penalized, minimum energy distributed control of the heat equation is considered. The problem may be stated as follows: minimize

$$J[u] = \alpha \int_0^{R_F} [x(r,\, T_f) - x_D(r)]^2 dr + \beta \int_0^{T_f} \int_0^{R_F} u^2(r,\, t)\,dr\,dt,$$

$$\alpha = \beta = 1,\ R_F = 1,\ T_f = 1$$

subject to

$$x_t = x_{rr} + u,\ x(r,\, 0) = 0,\ x(0,\, t) = 0,\ x(1,\, t) = 0,\ x_D(r) = r(1 - r).$$

The adjoint system of the above problem is given by

$$p_t = -p_{rr},\ p(r,\, T_f) = 2[x(r,\, T_f) - x_D(r)],\ p(0,\, t) = p(1,\, t) = 0,$$

with the gradient $g(r,\, t) = 2u(r,\, t) + p(r,\, t)$.

The results of this problem's solution by the conjugate gradient method are given in Table I.

Example 2. Distributed Control of the Vibrating String (unconstrained).

The unconstrained, fixed time, penalized, minimum energy distributed control of the vibrating string is considered. The problem may be stated as follows: minimize

$$J[u] = \alpha \int_0^{R_f} x^2(r,\, T_f)\,dr + \beta \int_0^{T_f} \int_0^{R_f} u^2(r,\, t)\,dr\,dt,\ \alpha = 5,\ \beta = 0.5,$$

$$R_F = 1,\ T_F = 4,$$

subject to

$$x_{tt} = x_{rr} + u,\ x(r,\ 0) = \sin \pi r,\ x_t(r,\ 0) = 0,\ x(0,\ t) = x(1,\ t) = 0,$$
$$0 \leq r \leq 1,\ 0 \leq t \leq 4.$$

The adjoint system for the above problem is given by

$$p_{tt} = p_{rr},\ p(r,\ T_f) = 0,\ p_t(r,\ T_f) = 10\,x(r,\ T_f),\ p(0,\ t) = p(1,\ t) = 0,$$

with the gradient $g(r,\ t) = u(r,\ t) + p(r,\ t)$.

The results of the solution by the conjugate gradient method as well as the comparison with the steepest descent solution are given in Table II.

Example 3. Distributed Control of the Vibrating String (constrained).

The fixed time, fixed terminal state (partial), minimum energy distributed control of the vibrating string is considered. The problem may be stated as follows: minimize

$$J[u] = \beta \int_0^{T_f} \int_0^{R_f} u^2(r,\ t)\,dr\,dt,\ \beta = 0.5,\ T_f = 4,\ R_f = 1$$

subject to

$$x_{tt} = x_{rr} + u,\ x(r,\ 0) = \sin \pi r,\ x(r,\ T_f) = 0,\ x_t(r,\ 0) = 0,$$
$$x(0,\ t) = x(1,\ t) = 0.$$

Introduction of the penalty function to account for the terminal constraint yields a new cost functional

$$J_p = \alpha_n \int_0^{R_f} x^2(r,\ T_f)\,dr + J[u],$$

where the penalty constant α_n is arbitrarily chosen. The adjoint system and the gradient of J_p are then exactly the same as in Example 2. The sequential unconstrained minimization procedure is to solve a sequence of unconstrained problems which converge to the solution of the constrained problem. Results of the solution with increasing α_n are given in Table III.

Example 4. Boundary Control of the Vibrating String

The fixed time, penalized, minimum energy boundary control of the vibrating string is considered. The problem may be stated as follows: minimize

$$J[u] = \alpha \int_0^{R_F} x^2(r,\ T_f)\,dr + \beta \int_0^{T_f} u^2(t)\,dr,\ \alpha = \beta = 1,\ R_f = 1,\ T_f = 4,$$

subject to

$$x_{tt} = x_{rr};\ x(r,\ 0) = \sin \pi r,\ x_t(r,\ 0) = 0,\ x(0,\ t) = u(t),\ x(1,\ t) = 0.$$

The gradient of the cost functional is given by $g(t) = u(t) + \left.\frac{\partial p}{\partial r}\right|_{r=0}$ where p is the adjoint variable defined by

$$p_{tt} = p_{rr},\ p(r,\ T_f) = 0,\ p_t(r,\ T_f) = 2x(r,\ T_f),\ p(0,\ t) = p(1,\ t) = 0.$$

The results of the solution of this problem by the conjugate gradient method are given in Table IV.

ACKNOWLEDGMENT

This work was supported in part by the Office of Naval Research under ONR Contract N00014-68-A-0162 and the Engineering Research Institute at Iowa State University.

REFERENCES

[1] Alexband, E. I., "Function Space Methods for the Optimal Control of a Class of Distributed Parameter Control Systems," Proc. JACC, 1965, Rensselear Polytechnic Institute, Troy, N.Y., pp. 374-380.

[2] Kim, M. M., "Successive Approximation Method in Optimum Distributed Parameter Systems," Journal of Optimization Theory and Applications, Vol. 4, No. 1, pp. 40-43, 1969.

[3] Sage, A. P., and Chaudhuri, S. P., "Gradient and Quasi-Linearization Computational Techniques for Distributed Parameter Systems," International Journal of Control, Vol. 6, No. 1, pp. 81-98, 1967.

[4] Denn, M. M., "Optimal Boundary Control for a Nonlinear Distributed System," International Journal of Control, Vol. 4, No. 2, pp. 167-178, 1966.

[5] Autosiewicz, H. A., and Rheinbolt, W. C., Conjugate-Direction Methods and the Method of Steepest Descent. In Todd, John, ed. Numerical Analysis and Functional Analysis, pp. 501-517, New York, N.Y., McGraw-Hill Book Co., Inc., 1962.

[6] Urabe, M., "Convergence of Numerical Iteration in Solution of Equations," Journal of Science of the Hiroshima University, Ser. A, Vol. 19, No. 3, January 1956.

[7] Cornick, D. E., "The Numerical Optimization of Distributed Parameter Systems by Gradient Methods," Ph.D. Dissertation. Ames, Iowa, Library, Iowa State University, 1970.

[8] Fletcher, R., and Powell, M. J. D., "A Rapidly Convergent Descent Method for Minimization," The Computer Journal, Vol. 6, pp. 163-168, 1963.

[9] Collatz, L., The Numerical Treatment of Differential Equations, Springer-Verlag, Berlin, 1960.

[10] Carnahan, B., Luther, H. A., Wilkes, J. O., Applied Numerical Methods, John Wiley and Sons, Inc., New York, N.Y., 1969.

Table I.[a] Results for Example 1.

Iteration number n	$\tilde{J}[\tilde{u}_n]$	$<\tilde{g}_n, \tilde{g}_n>$
0	0.78591E1	0.31317E2
1	0.34088E-1	0.53149E-3
2	0.33948E-1	0.10632E-6
3	0.33948E-1	0.32117E-12
4	0.33948E-1	0.10157E-15

Table III.[a] Results for Example 3.

Penalty constant W	$J = J_p - P[x]$	Constraint error
2	0.24954E0	0.14859E0
5	0.55010E0	0.52410E-1
50	0.10938E1	0.10421E-2
100	0.11454E1	0.27283E-3
500	0.11894E1	0.11353E-4
1000	0.11957E1	0.28465E-5

[a] All computations were performed in double precision. Only the first five significant digits are tabulated.

Table II.[a] Results of Example 2.

Iteration number n	Conjugate gradient method $\tilde{J}[\tilde{u}_n]$	$\langle \tilde{g}_n, \tilde{g}_n \rangle$	Steepest descent method $\tilde{J}[\tilde{u}_n]$	$\langle \tilde{g}_n, \tilde{g}_n \rangle$
0	0.82497E1	0.17787E2	0.82497E1	0.17787E2
1	0.23669E1	0.63747E1	0.23669E1	0.63747E1
2	0.81218E0	0.12307E-3	0.11382E1	0.77325E0
3	0.81214E0	0.25974E-4	0.88149E0	0.27711E0
4	0.81215E0	0.15799E-7	0.82691E0	0.33613E-1
5	0.81215E0	0.24469E-9	0.81544E0	0.12046E-1
6	0.81215E0	0.19422E-11	0.81292E0	0.14612E-2
7	0.81215E0	0.40834E-15	0.81235E0	0.52366E-3
Error Analysis (for conjugate gradient method only). Optimal Control Error: $\|\tilde{u}^* - u^*\| \leq 0.11593165E1$; Cost Functional Error: $\lvert J[u^*] - \tilde{J}[\tilde{u}^*]\rvert \leq 0.10543401E2$, where $\|e_g(u^*; h)\| \cong 0.11593165E1$, $\|\tilde{g}(\tilde{u}^*; qh)\|^2 = 0.75600837E0$, $\|\tilde{u}^*\| = 0.10489104E1$. CPU Time = 10 sec. (50 sec with error analysis). Storage = 29384 BYTES (56936 BYTES with error analysis)				

Table IV.[a] Results for Example 4.

Iteration number n	$\tilde{J}[\tilde{u}_n]$	$\langle \tilde{g}(\tilde{u}_n), \tilde{g}(\tilde{u}_n) \rangle$	Error Analysis. Optimal Control Error: $\|\tilde{u}^* - u^*\| \leq 0.17718170E0$; Cost Functional Error: $\lvert J[u^*] - \tilde{J}[\tilde{u}^*]\rvert \leq 0.20936213E0$, where $\|e_g(\tilde{u}^*; h)\| \cong 0.35436341E0$, $\|\tilde{g}(\tilde{u}^*; qh)\|^2 = 0.51743680E-1$, $\|\tilde{u}^*\| = 0.28581078E0$. CPU TIME = 10.58 sec. Storage = 32400 BYTES.
0	0.50185E0	0.88209E1	
1[b]	0.10027E0	0.31521E-28	

[a] All computations were performed in double precision. Only the first five significant digits are tabulated.

[b] Convergence in one iteration occurred only when u = 0 was used as the initial control guess.

*D. E. Cornick is presently with Martin-Marietta, Littleton, Colorado.

ITERATIVE MINIMIZATION IN HILBERT SPACE OF QUADRATIC FUNCTIONALS WITH NONUNIQUE MINIMA

M. VIDYASAGAR
Department of Electrical Engineering
Sir George Williams University
Montreal, Canada

ABSTRACT

It is shown that both the steepest descent and the conjugate gradient method can be used to iteratively minimize a class of quadratic functionals in a Hilbert space, even when the quadratic functional is minimized at an infinite number of points. Regardless of the starting point of the iteration, both the steepest descent and the conjugate gradient methods will always converge to an optimum (and in fact the same optimum), but two different iterations starting from two different points might converge to two different optima.

Consider a linear control system characterized by an operational relation of the form

$$q = H u \tag{1}$$

where q denotes the state of the system, u denotes the control function, and H denotes the system operator. The control u is a member of a Hilbert space X, termed the control space; the state q is a member of another Hilbert space Y, termed the state space; and the system operator H is a continuous linear mapping from X into Y. Suppose a "desired state" $q_o \in Y$ is specified, and consider the problem of finding a $u \in X$ which minimizes the functional

$$f(u) = (1/2) \| q_o - H u \|^2 \tag{2}$$

In other words, the objective is to minimize the error between the desired state q_o and the actual state $H\,u$.

Let $(.,.)$ denote the inner product on both X and Y, and let H' denote the adjoint operator of H, mapping Y into X. Then, the problem at hand can be restated as follows: Find a $u \in X$ which minimizes

$$f(u) = (1/2)\,(q_o - H\,u,\ q_o - H\,u)$$

$$= (1/2)\,(u,\ H'H\,u) - (u,\ H'\,q_o) + (1/2)\,(q_o, q_o) \qquad (3)$$

Alternately, if a functional $f(.)$ of the form (3) is given and one is required to minimize it over a Hilbert space X, such a problem can be interpreted in terms of the foregoing optimal control problem. Define

$$A = H'H;\ w = H'\,q_o;\ f_o = (q_o, q_o)/2 \qquad (4)$$

Then $A: X \to X$, $w \in X$, and

$$f = (1/2)\,(u,\ A\,u) - (u,\ w) + f_o \qquad (5)$$

In Ref. 1, Horowitz and Sarachik show that if A is positive definite, then there is exactly one $u \in X$ which minimizes $f(.)$. They also show that any one of three methods - namely the steepest descent method, and conjugate gradient method, and Davidon's method - can be used to iteratively compute this minimizing element. In Ref. 2, the case of a convex functional, but again with a unique minimum, is considered. However, no one to date appears to have considered the case where A is not necessarily definite (it is obviously positive). In such a case, no optimal control need exist, or there might be an infinite number of optimal controls.

In this paper, we tackle precisely such problems, and show that a minimizing element for $f(.)$ exists provided the operator

A is compact. It is further shown that if A is compact, both the steepest descent and the conjugate gradient methods will always converge to a minimizing element of f(.). If started from the same point, both methods give the same minimum, but two different iterations starting from two different points might converge to two different minima. A method is indicated for iteratively computing the minimizing element with the smallest norm.

Proposition 1

If A is compact (i.e. maps closed and bounded sets into compact sets), then there is an element in X which minimizes f(.).

Proof

It is very easy to show that (Ref. 3, p. 161) the elements in X (if any) which minimize f(.) are precisely the solutions of the equation

$$H'H\ u = H'\ q_o \tag{6}$$

Suppose A = H'H is compact. Then (Ref. 4, p. 123) a necessary and sufficient condition for the above equation to have a solution is that $H'\ q_o$ is orthogonal to every solution of the homogeneous equation $H'H\ u = 0$; i.e. $(H'\ q_o,\ u) = 0$ whenever $H'H\ u = 0$. We shall show that this latter condition is always satisfied, regardless of q_o. Suppose $H'H\ u = 0$. Then $(H\ u,\ H\ u) = (u,\ H'H\ u) = 0$, whence $H\ u = 0$. Therefore $(H'\ q_o,\ u) = (q_o,\ H\ u) = 0$, and it follows that (6) always has a solution, whatever q_o might be.

Proposition 2

If A is compact, then H(X) is a closed subspace of Y.

Proof

If $H(X) = Y$, then by definition it is a closed subspace, of Y. So we only need to consider the case $H(X) \neq Y$. Suppose by way of contradiction that $H(X)$ is not closed in Y, and let $q_1 \varepsilon Y$ be chosen such that $q_1 \varepsilon \overline{H(X)}$ (the closure of $H(X)$), but $q_1 \not\varepsilon H(X)$. Define $\phi(u) = \| q_1 - H u\|^2$. Then clearly $\inf_{u \varepsilon X} \phi(u) = 0$, but no u exists that $\phi(u) = 0$. However, this contradicts the results of Proposition 1, which (when slightly modified), state that $\phi(.)$ has a minimizing element in X. Hence no such q_1 exists, and the proposition follows.

Proposition 3

Let N denote the null space of A, i.e.

$$N = \{u \varepsilon X : A u = 0\} \tag{7}$$

If A is compact, then A is positive definite on X/N (the quotient space of X with respect to N).

Proof

It is clear from the definition of A that it is positive. Also, A is one-to-one when viewed as a mapping from X/N into itself. Hence $A(X/N)$ is at least dense in X/N. By a slight modification of Proposition 2, it follows that $A(X/N)$ is closed in X/N. Thus, by (Ref. 5, Thm. II.2.2) it follows that A^{-1} is continuous, if it exists. All this together shows that A is positive definite on X/N.

This proposition suggests a means for determining a minimizing element of $f(.)$ in cases where A is not definite. First, let M denote the orthogonal complement (in X) of N. Then it is clear that X/N is isomorphic to M, so that A is positive definite on M. Let P denote the projection of X onto M. Then for any u in X, the corresponding element in X/N is (in the sense of the isomorphism between X/N and M) equal to P u. To apply the iteration schemes, pick u_1 in X arbitrarily; then P u_1 is the starting point for the iterations in X/N. The next step in the iteration, say u_2, does not necessarily belong to X/N, in which case, to apply the steepest descent or conjugate gradient schemes, the next step must be chosen as P u_2. If one proceeds thus, it follows from the results of Ref. 1 that $\{P\ u_n\}$ converges to P u_o, where u_o is some minimizing element of f(.). In fact, it is not difficult to show that the set of all minimizing elements of f(.) is given by

$$S = \{u + P\ u_o : u \varepsilon N\} \tag{8}$$

While the above iteration scheme is theoretically possible, it is not very feasible in practice because it involves explicitly determining the null space N and the projection operator P, both of which are very complicated tasks. In what follows, we shall prove two simple relations which in effect make it unnecessary to compute N or P or any related entity. In fact, one can "blindly" apply either the steepest descent or the conjugate gradient method without knowing beforehand whether A is positive definite or not, and still the iterations will converge to an optimum. Before proceeding further, the details of the two methods of iteration are recalled in the interests of convenience

In both methods of iteration, a starting point u_1 is chosen arbitrarily. Subsequent iterations are determined by the rules

$$u_{n+1} = u_n + c_n\ p_n \tag{9}$$

$$c_n = (p_n,\ g_n)/(p_n,\ A\ p_n) \tag{10}$$

$$g_n = A\ u_n - w \tag{11}$$

Note that g_n is the Fréchet derivative of $f(.)$ at u_n. In the steepest descent method, the "direction of iteration" p_n is chosen as $- g_n$, while in the conjugate gradient method, p_n is chosen as

$$p_1 = -g_1 ; \; p_n = -g_n + (g_n, g_n)\, g_{n-1}/(g_{n-1}, g_{n-1}) \quad \text{for all } n \geq 1 \tag{12}$$

Let $g(u) = A\,u - w$ denote the Fréchet derivative of $f(.)$ at u (note that g_n is just a shorthand notation for $g(u_n)$).

Proposition 4

$g(P\,u) = g(u)$ for all u in X.

Proof

We have $u = P\,u + (I - P)u$, where I is the identity operator, and $g(u) = g(P\,u) + A(I - P)\,u$. But $(I - P)\,u \in N$ for all u, so that $A(I - P)\,u = 0$, and the result follows.

Proposition 5

$P\,g(u) = g(u)$ for all u in X.

Proof

We have $g(u) = H'H\,u - H'\,q_o = H'(H\,u - q_o)$. Hence the result follows if we show that the range of H' is contained in M. From Proposition 2, we know that $H(X)$ is closed in Y. Hence from (Ref. 5, Thm. VI.6.2),

$$\overline{H'(Y)} = \{x \in X : (x,y) = 0 \text{ whenever } H\,y = 0\} \tag{13}$$

clearly $\overline{H'(Y)} = M$, whence $H'(Y)$ is contained in M, and the Proposition is proved.

Proposition 6

With the u_n defined by (10) - (12),

$$u_n = P\,u_n + (I - P)\,u_1 \quad \text{for all } n \geq 1 \tag{14}$$

Proof

In the case of the steepest descent method, (14) is obvious by induction. In the case of the conjugate gradient method: since p_n is a linear combination of $g_1, \ldots, g_n$, Propositions 4 and 5 hold with g(.) replaced by p_n. Hence the above result is valid regardless of which method is used to generate the sequence $\{u_n\}$.

We are now in a position to state the final results of the paper. First, it is clear from Proposition 5 thru 6 that iterating in X amounts to the same thing as iterating in X/N, except for the additive term $(I - P)\ u_1$ which appears at each step of the former. Secondly, from previous discussion, it follows that $\{P\ u_n\}$ converges to $P\ u_o$, where u_o is any minimizing element of f(.). Hence, we conclude from Proposition 6 that the sequence of iterations $\{u_n\}$ converges to $P\ u_o + (I - P)\ u_1$. Various other results follow immediately: 1) Both the steepest descent and conjugate gradient methods converge to a minimizing element of f(.), regardless of the starting point. 2) If started from the same point, both methods converge to the same point. 3) If two iterations are started from two different points u_1 and v_1, and if the projections of u_1 and v_1 onto N are equal, then both iterations converge to the same point. 4) In cases where there are nonunique minima, it is often of interest to compute the minimizing element with the smallest norm (it is easy to show that such an element exists). To iteratively compute such an element, it suffices to pick u_1 such that $(I - P)\ u_1 = 0$, i.e. such that $u_1 \varepsilon M$. The simplest means of doing this is to choose $u_1 = H'\ y$, y being any element in Y. Then by Proposition 5, $u_1 \varepsilon M$. If u_1 is chosen in the fashion, the limit of the resulting iteration will be a) the unique minimizing element of f(.) if A is positive definite, or b) the minimizing element of f(.) with the smallest norm if A is not definite. It is once again emphasized that the contribution of this paper is in demonstrating that the steepest descent and conjugate gradient methods can be used to iteratively find a minimum without knowing beforehand whether A is definite or not.

Finally, the assumption of compactness on A is discussed briefly. While this is a fairly severe assumption mathematically, from a physical point of view this assumption is not particularly restrictive. For instance, many distributed systems are characterized by control-state relations of the form

$$q(x) = \int_0^t h(x,\tau)\; u(\tau)\; d\tau \;, \; o \leq x \leq a \qquad (15)$$

where the control u(.) belongs to $L^2[0,t]$, the state q(.) belongs to $L^2[0,a]$, and h(.,.) is a Fredholm kernel over the set [0,a] x [0,t]. An example of a system of the form (15) is the diffusion system in (Ref. 6, Example 1). For systems of the form (15), it is a well-known result in the theory of integral equations (Ref. 7, p. 302) that the operator H'H is compact.

REFERENCES

1. Horowitz, L.B., and Sarachik, P.E., Davidon's Method in Hilbert Space, SIAM Journal on Applied Mathematics, vol. 10, no. 3, pp. 676-695, 1968.

2. Daniel, J.W., The Conjugate Gradient Method for Linear and Nonlinear Operator Equations, Ph.D. Dissertation, Stanford University, Stanford, California, 1965.

3. Luenberger, D.G., Optimization by Vector Space Methods, John Wiley and Sons, New York, N.Y., 1969.

4. Mikhlin, S.G., Linear Integral Equations, Hindustan Publishers,Delhi, India, 1960.

5. Dunford, N., and Schwartz, J.T., Linear Operators, Part I, Interscience Publishers, New York, N.Y., 1959.

6. Vidyasagar, M., and Higgins, T.J., A Controllability Criterion for Stable Linear Systems, IEEE Transactions on Automatic Control, vol. AC-15, no. 3, pp. 391-392, 1970.

7. Kantorovich, L.V., and Akilov, G.P., Functional Analysis in Normed Spaces, Pergamon Press, Oxford, England, 1964.

OPTIMAL CONTROL WITH A BOUNDED ENERGY INPUT DERIVATIVE FOR A CLASS OF DISTRIBUTED SYSTEMS

YAAKOV YAVIN
Department of Electrical Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Vriginia 24061

ABSTRACT

An optimal control problem, for a class of distributed parameter systems, with a constraint on the input derivative, is posed. The constraint is a bound on the L_2 norm of the input derivative. Existence of optimal controls is proved, and a nonlinear integrodifferential equation is derived to serve as a necessary and sufficient condition for the optimal controls to satisfy. Also, a computational scheme, for optimal controls is suggested.

I. Introduction

Let a one-dimensional linear distributed parameter system be given by:

$$Q(x,t;u) = Q_o(x,t) + \int_o^t h(x,t,s)u(s)ds \quad , \quad 0 \leq x \leq 1, \quad 0 \leq t \leq T, \tag{1}$$

where Q_o and h are given functions, and $\frac{\partial Q_o(x,t)}{\partial x} \varepsilon L_2(0,1)$ for any $t \varepsilon [0,T]$, where $L_2(0,1) \triangleq \{Q \mid \int_o^1 [Q(x)]^2 dx < \infty\}$, (the integral in the sense of Lebesgue is used in this paper); and $\frac{\partial h(x,t,s)}{\partial x}$ is a bounded sectionally continuous function in $\{0 \leq x \leq 1, \ 0 \leq t, \ s \leq T\}$. $Q(x,t;u)$ is the system's response, and u is the control function.

Given the set

$$\Omega \triangleq \{ u \mid [\int_o^T (\frac{du(t)}{dt})^2 dt]^{1/2} \leq \alpha, \ |u(T)| \leq \beta\}$$

where α and β are given positive numbers; the following problem is posed:

Given a function $Q_d(x)$, $\frac{dQ_d}{dx} \varepsilon L_2(0,1)$, choose a control $u^* \varepsilon \Omega$ such that:

$I(u) \geq I(u^*)$ for any $u \varepsilon \Omega$, where

$$I(u) = \int_o^1 [Q_d(x) - Q(x,T;u)]^2 dx + \int_o^1 [\frac{dQ_d(x)}{dx} - \frac{\partial Q(x,T;u)}{\partial x}]^2 dx, \tag{2}$$

and T is a given positive number.

A control $u^* \varepsilon \Omega$ such that $I(u) \geq I(u^*)$ for all $u \varepsilon \Omega$ will be called an optimal control.

The problem treated in this paper can be considered as a special case of a more general optimal control problem in distributed parameter systems posed by Butkovskii and Lerner (1960). A considerable amount of work has been done in the area of optimal control of distributed parameter systems (see Robinson, (1969), for an almost complete list of references) especially for the case where the inputs are subjected to constraints. For example see: Butkovskii (1961), Balaksrishnan (1965), Sakawa (1966), Axelband (1966), Yavin and Sivan (1968), Yavin (1969), Weigand and Dsouza (1969), Vidyasager (1969). However, the optimal control problem as posed by Butkovskii and Lerner (1960), where the inputs derivatives are subjected to constraints, has not been considered.

This paper proves the existence of optimal controls to the problem posed here, and provides a necessary and sufficient condition for them in the form of a nonlinear integrodifferential equation. Also, a computational scheme, for optimal controls, in the case of $\beta = 0$, is suggested.

2. Existence of Optimal Controls

Lemma 1

The set Ω is a compact set of $C[0,T]$. ($C[0,T]$ is the Banach space of all real-valued, bounded continuous functions $f(t)$ defined on $[0,T]$, with $||f|| = \max\limits_{0 \leq t \leq T} |f(t)|$.)

Proof

Let $u \varepsilon \Omega$, then $u(t)$ can be written in the form

$$u(t) = -\int_t^T \frac{du(\tau)}{d\tau} d\tau + u(T). \tag{3}$$

For any $u \varepsilon \Omega$, by using equation (3) and Cauchy-Schwarz inequality, Liusternik and Sobolev (1961), the following inequalities are obtained:

$$\begin{aligned} |u(t)| &\leq \left|\int_t^T \frac{du(\tau)}{d\tau} d\tau\right| + |u(T)| \\ &\leq \left[\int_t^T d\tau\right]^{1/2} \left[\int_t^T \left|\frac{du(\tau)}{d\tau}\right|^2 d\tau\right]^{1/2} + |u(T)| \\ &\leq T\alpha + \beta \overset{\Delta}{=} M. \end{aligned} \tag{4}$$

Inequality (4) implies that, for any $u \varepsilon \Omega$

$$\int_o^T [u(t)]^2 dt + \int_o^T \left[\frac{du(t)}{dt}\right]^2 dt \leq M^2 T + \alpha^2 . \tag{5}$$

Using this result, it follows from Mikhlin and Smolitsky (1967), pp. 154, that Ω is a compact set of $C[0,T]$.

Theorem 1

There always exists a control $u^* \epsilon \Omega$, such that $I(u) \geq I(u^*)$ for any $u \epsilon \Omega$.

Proof

The proof follows from the fact that $I(u)$ is a continuous functional on the compact set Ω.

3. Conditions Upon the Optimal Controls

For convenience the following notations will be used:

$$Q_D(x) \overset{\Delta}{=} Q_d(x) - Q_o(x,T) \quad , \quad 0 \leq x \leq 1 \quad , \tag{6}$$

$$f(t) \overset{\Delta}{=} \int_o^1 h(x,T,t) Q_D(x) dx \quad , \quad 0 \leq t \leq T \quad , \tag{7}$$

$$g(t) \overset{\Delta}{=} \int_o^1 \frac{\partial Q_D(x)}{\partial x} \; \frac{\partial h(x,T,t)}{\partial x} dx \quad , \quad 0 \leq t \leq T \quad , \tag{8}$$

$$\phi(t) \overset{\Delta}{=} f(t) + g(t) \quad , \quad 0 \leq t \leq T \quad , \tag{9}$$

$$R(t,s) \overset{\Delta}{=} \int_o^1 h(x,T,t) h(x,T,s) dx \quad , \quad 0 \leq t,s \leq T \quad , \tag{10}$$

$$F(t,s) \overset{\Delta}{=} \int_o^1 \frac{\partial h(x,T,t)}{\partial x} \; \frac{\partial h(x,T,s)}{\partial x} dx \quad , \quad 0 \leq t,s \leq T \quad , \tag{11}$$

$$K(t,s) \overset{\Delta}{=} R(t,s) + F(t,s) \quad , \quad 0 \leq t,s \leq T. \tag{12}$$

Theorem 2

If $I(u^*) = 0$ for $u^* \epsilon \Omega$, then u^* is a solution to

$$\phi(t) - \int_o^T K(t,s) u(s) ds = 0 \quad , \quad 0 \leq t \leq T. \tag{13}$$

Proof

Let $I(u^*) = 0$ for $u^* \epsilon \Omega$. Then, from equation (2) it follows that u^* is a solution to equations (14)-(15)

$$Q_D(x) - \int_o^T h(x,T,s) u(s) ds = 0 \quad , \quad 0 \leq x \leq 1 \quad , \tag{14}$$

$$\frac{\partial Q_D(x)}{\partial x} - \int_0^T \frac{\partial h(x,T,s)}{\partial x} u(s)ds = 0 \quad , \quad 0 \le x \le 1. \tag{15}$$

From the multiplication of both sides of (14) with $\int_0^1 h(x,T,t)(\cdot)dx$, and both sides of (15) with $\int_0^1 \frac{\partial h(x,T,t)}{\partial x}(\cdot)dx$, and addition of the results equation (13) follows immediately.

Theorem 3

For a control $u^*\varepsilon\Omega$, to satisfy: $I(u) \geq I(u^*) > 0$, for any $u\varepsilon\Omega$, it is necessary and sufficient that it be a solution to

$$\frac{du(t)}{dt} = -\alpha \frac{\int_0^t [\phi(\tau) - \int_0^T K(\tau,s)u(s)ds]d\tau}{[\int_0^T \{\int_0^\tau [\phi(\lambda) - \int_0^T K(\lambda, s)u(s)ds]d\lambda\}^2 d\tau]^{1/2}} \quad , \tag{16}$$

$0 \le t \le T$,

$$u(T) = \beta \operatorname{sign}\{\int_0^T [\phi(t) - \int_0^T K(t,s)u(s)ds]dt\} \quad ,$$

where $\operatorname{sign} C \triangleq \frac{C}{|C|}$.

Proof

The proof is based on the following lemma.

Lemma 2

Let B_1 and B_2 be normed linear spaces. Let $||\cdot||_1$ denote the norm on B_1, and $||\cdot||_2$ denote the norm on B_2. Let B_2^* be the conjugate space of B_2, and let $||\cdot||$ denote the norm on that space. Let U be a convex set of B_1. Assume that S is a continuous linear mapping from U into B_2.

Let there be given an element Q_D, $Q_D \varepsilon B_2$. Denote $J(u) \triangleq ||Q_D - Su||_2$. Then, for an element $u^*\varepsilon U$ to satisfy $J(u) \geq J(u^*) > 0$ for any $u\varepsilon U$ it is necessary and sufficient that there be a linear functional $L\varepsilon B_2^*$ such that

$$\text{I.} \quad ||L|| = 1 \tag{17}$$

$$\text{II.} \quad L(Su^*) = \sup_{u\varepsilon U} L(Su) \tag{18}$$

$$\text{III.} \quad L(Q_D - Su^*) = ||Q_D - Su^*||_2. \tag{19}$$

Proof

This lemma is an immediate result of Theorem 2.5 of Deutsch and Maserik (1967).

For the proof of Theorem 3, let $B_1 = C[0,T]$, and $U = \Omega$ and $B_2 = W^{1,2}(0,1)$. (For details on Sobolev Spaces $W^{k,r}(0,1)$ see for example Yosida, (1968), pp. 55.) From these it follows that $I(u) = ||Q_D - Su||_2^2$, equation (2), where the linear operator S from U into $W^{1,2}(0,1)$ is given by

$$Su \triangleq \int_0^T h(\cdot,T,s)u(s)ds \quad . \tag{20}$$

Now, since $W^{1,2}(0,1)$ is an Hilbert space, every linear continuous functional L on $W^{1,2}(0,1)$ can be written as

$$L(Q) = \int_0^1 \psi(x)Q(x)dx + \int_0^1 \frac{d\psi(x)}{dx}\frac{dQ(x)}{dx}dx \quad , \tag{21}$$

where $Q\varepsilon W^{1,2}(0,1)$ and $\psi\varepsilon W^{1,2}(0,1)$.

Thus, the operation of L on Su, (20), yields:

$$L(Su) = \int_0^1\{\psi(x)\int_0^T h(x,T,t)u(t)dt + \frac{d\psi(x)}{dx}\int_0^T \frac{\partial h(x,T,t)}{\partial x}u(t)dt\}dx \tag{22}$$

$$= \int_0^T H(t)u(t)dt$$

where

$$H(t) \triangleq \int_0^1[\psi(x)h(x,T,t) + \frac{d\psi(x)}{dx}\frac{\partial h(x,T,t)}{\partial x}]dx, \quad 0 \leq t \leq T. \tag{23}$$

The insertion of equation (3) into equation (22) yields

$$L(Su) = -\int_0^T(\int_0^t H(\tau)d\tau)\frac{du(t)}{dt}dt + u(T)\int_0^T H(t)dt. \tag{24}$$

Condition II of Lemma 2 is satisfied if and only if u* is a solution to

$$\frac{du(t)}{dt} = -\alpha\frac{\int_0^t H(\tau)d\tau}{[\int_0^T\{\int_0^t H(\tau)d\tau\}^2dt]^{1/2}} \quad , \qquad 0 \leq t \leq T, \tag{25}$$

$$u(T) = \beta\, \text{sign}\{\int_0^T H(t)dt\} \quad .$$

Conditions I and III of Lemma 2 are satisfied if and only if ψ is given by

$$\psi(x) = \frac{1}{C}[Q_D(x) - \int_0^T h(x,T,t)u^*(t)dt] \quad , \tag{26}$$

where

$$C = ||Q_D - Su^*||_2 =$$

$$= [\int_o^1 \{Q_D(x) - \int_o^T h(x,T,s)u^*(s)ds\}^2 dx + \int_o^1 \{\frac{\partial Q_D(x)}{\partial x} - \int_o^T \frac{\partial h(x,T,s)}{\partial x} u^*(s)ds\}^2 dx]^{1/2} \tag{27}$$

If ψ and $\frac{d\psi}{dx}$, (26), are substituted into H(t), (23), and the notations (7)-(12) are used, then equation (25) implies equation (16). Thus, equation (16) turns out to be a necessary and sufficient condition for optimality.

4. Computation of Optimal Controls

In this section a method for solving equation (16) for the case $\beta = 0$ is suggested.

Assume $\beta = 0$. Then, using equation (3), equation (16) becomes

$$\frac{du(t)}{dt} = -\alpha \frac{\int_o^t [\phi(\tau) + \int_o^T \int_o^z K(\tau,s)ds \frac{du(z)}{dz} dz] d\tau}{[\int_o^T \{ \int_o^t [\phi(\tau) + \int_o^T \int_o^z K(\tau,s)ds \frac{du(z)}{dz} dz] d\tau\}^2 dt]^{1/2}} \tag{28}$$

$0 \leq t \leq T$

$u(T) = 0.$

Let h(x,T,t) and $Q_D(x)$, (6), be given by:

$$h(x,T,t) = \sum_{n=1}^{\infty} W_n(t) \sin n\pi x \quad \text{a.e. in } 0 \leq x \leq 1, \quad 0 \leq t \leq T \ , \tag{29}$$

$$Q_D(x) = \sum_{n=1}^{\infty} q_n \sin n\pi x \quad \text{a.e. in } 0 \leq x \leq 1 \ , \tag{30}$$

where $\{W_n(t)\}_{n=1}^N$ is a system of linearly independent, continuous functions for every integer N.

By the insertion of (29)-(30) into (7)-(12) the following expressions are obtained:

$$\phi(t) = \frac{1}{2} \sum_{n=1}^{\infty} \omega_n q_n W_n(t) \ , \qquad 0 \leq t \leq T \ , \tag{31}$$

$$K(t,s) = \frac{1}{2} \sum_{n=1}^{\infty} \omega_n W_n(t) W_n(s) \ , \qquad 0 \leq t,s \leq T \ , \tag{32}$$

where

$$\omega_n \stackrel{\Delta}{=} 1 + n^2\pi^2 \ , \qquad n = 1,2,... \ . \tag{33}$$

Define

$$\phi_N(t) \triangleq \frac{1}{2} \sum_{n=1}^{N} \omega_n q_n W_n(t) \quad , \qquad 0 \leq t \leq T \ , \tag{34}$$

$$K_N(t,s) \triangleq \frac{1}{2} \sum_{n=1}^{N} \omega_n W_n(t) W_n(s) \quad , \qquad 0 \leq t,s \leq T, \tag{35}$$

$$G_n(t) \triangleq \int_o^t W_n(\tau) d\tau, \qquad 0 \leq t \leq T \ , \quad n = 1,2,\ldots \ . \tag{36}$$

Remark: $\{G_n(t)\}_{n=1}^{N}$ is a system of linearly independent functions for every integer N; this is implied from the linear independence of the system $\{W_n(t)\}_{n=1}^{N}$.

By the insertion of (34)-(35) into (28), an equation for an approximate optimal control, is obtained:

$$\frac{du(t)}{dt} = -\alpha \frac{\sum_{n=1}^{N} \omega_n (q_n + v_n) G_n(t)}{\left[\sum_{n=1}^{N} \sum_{m=1}^{N} \omega_n \omega_m (q_n + v_n)(q_m + v_m) G_{nm}\right]^{1/2}} \quad , \tag{37}$$

$$0 \leq t \leq T \ ,$$

$$u(T) = 0 \quad ,$$

where

$$v_n \triangleq \int_o^T G_n(t) \frac{du(t)}{dt} \, dt \quad , \qquad n = 1,2,\ldots \quad , \tag{38}$$

and

$$G_{nm} \triangleq \int_o^T G_n(t) G_m(t) dt \quad , \qquad n,m = 1,2,\ldots \ . \tag{39}$$

The solution to (37) will be denoted by $u_N(t)$ and will serve as an approximation to an optimal control.

Once the set of numbers $\{v_n\}_{n=1}^{N}$ is found, then it follows from (37), that $u_N(t)$ is given by

$$u_N(t) = \frac{\alpha}{\lambda_N} \sum_{n=1}^{N} \omega_n (q_n + v_n) \int_t^T G_n(\tau) d\tau \qquad , \quad 0 \leq t \leq T \quad , \tag{40}$$

where

$$\lambda_N \triangleq [(\hat{q} + \hat{v})^T AGA (\hat{q} + \hat{v})]^{1/2} \qquad , \tag{41}$$

and

$$\hat{q} \triangleq \text{col. } (q_1, q_2, \dots, q_N) \quad , \tag{42}$$

$$\hat{v} \triangleq \text{col. } (v_1, v_2, \dots, v_N) \quad , \tag{43}$$

$$A \triangleq \begin{pmatrix} \omega_1 & 0 & 0 & \dots & 0 \\ 0 & \omega_2 & 0 & \dots & 0 \\ 0 & 0 & \omega_3 & \dots & 0 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 0 & 0 & 0 & \dots & \omega_4 \end{pmatrix} , \tag{44}$$

$$G \triangleq \begin{pmatrix} G_{11} & G_{12} & \dots\dots & G_{1N} \\ G_{21} & G_{22} & \dots\dots & G_{2N} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ G_{N1} & G_{N2} & \dots\dots & G_{NN} \end{pmatrix} . \tag{45}$$

Multiplying both sides of (37) by $\int_o^T G_m(t)\,(\cdot)dt$, for m = 1,2,...,N, yields:

$$\hat{v} = -\frac{\alpha}{\lambda_N}\, GA(\hat{q} + \hat{v}) \quad . \tag{46}$$

Using the notation

$$\gamma \triangleq \lambda_N/\alpha \tag{47}$$

and solving (46) for $\hat{v}$ gives

$$\hat{v} = -(\gamma I + GA)^{-1} GA\hat{q} \tag{48}$$

where I is the N-dimensional unit matrix.

By inserting (48) into (41), the following equation for γ is obtained:

$$\gamma^2\alpha^2 = \hat{q}^T[I-(\gamma I+GA)^{-1}GA]^T AGA[I-(\gamma I+GA)^{-1}GA]\hat{q} \quad . \tag{49}$$

Equation (49) can be reduced to a polynomial equation $P_M(\gamma) = 0$, where $P_M(\gamma)$ is a polynomial in γ of degree M = 2N + 2.

The computational scheme for $u_N(t)$ may be summarized as follows:

(a) Find a positive root, γ, of equation (49).

(b) Compute $\hat{v}$ from equation (48), where the positive root of (49) is used for γ.

(c) Compute $u_N(t)$ from (40), by using γ of (a), and $\hat{v}$ of (b).

For the validity and applicability of such methods see Yavin and Sivan (1968), and Yavin and Rasis (1970).

REFERENCES

Axelband, E.I., 1966, I.E.E.E. Trans. Autom. Control, 11, 42.

Balakrishnan A.V., 1965, J. SIAM Control, A, 3, 152.

Butkovskii A.G. and Lerner A.Y. 1960, Autom. and Remote Control, 21, 472.

Butkovskii A.G., 1961, Autom. and Remote Control, 22, 13.

Deutsch R.F. and Maserick P.H., 1967, SIAM Review, 9, 516.

Liusternik L.A. and Sobolev V.J., 1961, "Elements of Functional Analysis", (New York: Frederick Ungar Publishing Company.)

Mikhlin S.G. and Smolitsky K.L., 1967, "Approximate Methods for Solution of Differential and Integral Equations", (New York: American Elsevier Publishing Company, Inc.)

Robinson, A.C., 1969, "A Survey of Optimal Control of Distributed-Parameter Systems", (Office of Aerospace Research, USAF, ARL 69-0177).

Sakawa Y., 1966, I.E.E.E. Trans. Autom. Control, 11, 35.

Vidyasagar M., 1969, Reprints for J.A.C.C., 10, 694, (Boulder, Colorado.)

Weigand W.A. and D'souza A.F., 1969, Trans. of the ASME, J. of Basic Eng., pp. 161-167.

Yavin Y. and Sivan R., 1968, Int. J. Control, 8, 525.

Yavin Y., 1969, J. of Computer and System Sciences, 3, 24.

Yavin Y. and Rasis Y., 1970, Int. J. Control, 11, 153.

Yosida K., 1968, "Functional Analysis", (Springer-Verlag New York Inc.).

BOUNDED ENERGY OPTIMAL CONTROL OF TRANSVERSE VIBRATIONS OF A BEAM

Y. YAVIN
Department of Electrical Engineering
Virginia Polytechnic Institute
Blacksburg, Virginia

Y. RASIS
Department of Electrical Engineering
Technion, Israel Institute of Technology
Haifa, Israel

ABSTRACT

An optimal control problem for the transverse vibrations of a beam, for three kinds of boundary conditions, is posed. A complete scheme for computing optimal control is suggested.

This scheme comprises two steps:

(a) approximation of the transverse displacements by the Method of Moments;

(b) approximation of the cost functional by a sequence of functionals;

The minimization procedure is applied to each of these functionals separately.

1. Introduction

Consider the system whose dynamic behaviour is described by:

$$\rho(x)A(x)\frac{\partial^2 u(x,t)}{\partial t^2}+\frac{\partial^2}{\partial x^2}\left(E(x)I(x)\frac{\partial^2 u(x,t)}{\partial x^2}\right)=r(x)f(t) \tag{1}$$

$$t \geq 0 \quad , \quad 0 \leq x \leq 1$$

where $\rho(x)$, $A(x)$, $E(x)$, $I(x)$ are given continuous functions, and $\rho(x) \geq \rho_o > 0$, $A(x) \geq A_o > 0$, $E(x) \geq E_o > 0$, $I(x) \geq I_o > 0$ for all $x \varepsilon [0,1]$, and $E, I \varepsilon C_2[0,1]$.*

---

* $C_k[0,1] \overset{\Delta}{=} \{f \mid f^{(m)}(x)$ are continuous functions for all $x \varepsilon [0,1]$, for $m = 0,1,\ldots,k\}$.

$r(x)$ is a given function, $r \in L_2(0,1)$.**

(1) is the equation of transverse vibrations of a beam where $u(x,t)$ is the transverse displacement and $f(t)$ is the control function. For further details see [1], where the physical meaning of (1) and of ρ, A, E, I are given.

$u(x,t)$ obeys one of the following three conditions [1]:

$$\left\{\begin{array}{l} u(0,t) = u(1,t) = 0 \\ \dfrac{\partial u(0,t)}{\partial x} = \dfrac{\partial u(1,t)}{\partial x} = 0 \end{array}\right. \tag{2}$$

(built-in ends);

$$\left\{\begin{array}{l} u(0,t) = u(1,t) = 0 \\ \left[E(x)I(x)\dfrac{\partial^2 u(x,t)}{\partial x^2}\right]_{x=0} = \left[E(x)I(x)\dfrac{\partial^2 u(x,t)}{\partial x^2}\right]_{x=1} = 0 \end{array}\right. \tag{3}$$

(free supported ends);

$$\left\{\begin{array}{l} u(0,t) = \dfrac{\partial u(0,t)}{\partial x} = 0 \\ \left[E(x)I(x)\dfrac{\partial^2 u(x,t)}{\partial x^2}\right]_{x=1} = \left[\dfrac{\partial}{\partial x}\left(E(x)I(x)\dfrac{\partial^2 u(x,t)}{\partial x^2}\right)\right]_{x=1} = 0 \end{array}\right. \tag{4}$$

(x=0 is a built-in end and x=1 is a free end).

The initial conditions are

$$u(x,0) = 0 \quad \frac{\partial u(x,0)}{\partial t} = 0 \ . \tag{5}$$

Denote by $u(x,t;f)$ the solution to (1) and (5) which obeys one of the three boundary conditions.

Given a function $u_d(x)$, $u_d \in L_2(0,1)$, the problem is to choose a control function $f \in \Omega$ where $\Omega \triangleq \{f \mid \int_0^T [f(t)]^2 dt \leq 1\}$, and T is a given number, so that the functional

$$I(f) = \int_0^1 [u_d(x) - u(x,T;f)]^2 dx \tag{6}$$

will be minimized.

** $L_2(0,1) \triangleq \{f \mid \int_0^1 [f(x)]^2 dx < \infty\}$.

The system described by (1) with boundary conditions (2) or (3) or the case of free ends is dealt with by Komkov [1]. In [1] optimal control problems are defined and existence and uniqueness of optimal controls is established. Also conditions for optimum are given. No algorithm for solution is given.

The purpose of this paper is to suggest a complete scheme for the solution of the problem posed here, for each of the sets of boundary conditions. This is done by using the Method of Moments for approximating $u(x,t;f)$, and then by approximating $I(f)$ by a sequence of functionals to each of which the minimization procedure is applied.

In the following section it will be shown that the system (1), for each of the sets of boundary conditions, and for any $f \varepsilon \Omega$, has a solution $u(x,t;f)$ and $u(\cdot,t;f) \varepsilon L_2(0,1)$.

Since Ω is a convex, bounded and closed set in $L_2(0,T)$, and $I(f)$ is a convex continuous functional in Ω, it follows from [2] that $I(f)$ assumes its minimum in Ω.

2. The Solution of the Partial Differential Equation

By introducing the differential operator

$$L \overset{\Delta}{=} \frac{1}{\rho(x)A(x)} \frac{\partial^2}{\partial x^2}\left(E(x)I(x)\frac{\partial^2}{\partial x^2}\right) , \tag{7}$$

(1) can be written in the form

$$\frac{\partial^2 u}{\partial t^2} = -Lu + Qf(t) , \tag{8}$$

where

$$Q(x) \overset{\Delta}{=} r(x)/(\rho(x)\cdot A(x)) . \tag{9}$$

Define the following sets:

$$D_a \overset{\Delta}{=} \{u \mid u^{(4)} \varepsilon L_2(0,1) , u(0) = u(1) = 0 \quad u^{(1)}(0) = u^{(1)}(1) = 0\} \tag{10}$$

$$D_b \overset{\Delta}{=} \{u \mid u^{(4)} \varepsilon L_2(0,1) , u(0) = u(1) = 0 \quad u^{(2)}(0) = u^{(2)}(1) = 0\} \tag{11}$$

$$D_c \overset{\Delta}{=} \{u \mid u^{(4)} \varepsilon L_2(0,1) , u(0) = u^{(1)}(0) = 0 \; u^{(2)}(1) = u^{(3)}(1) = 0\}. \tag{12}$$

Using [3] it can be shown that each of the sets D_a, D_b, and D_c is a dense set in $L_2(0,1)$.

The operation L defined over each of the sets D_a, D_b, and D_c, defines three linear operators which will be denoted here by L_a, L_b, and L_c respectively.

By introducing the inner product $\langle u,v\rangle \triangleq \int_0^1 \rho(x)A(x)u(x)v(x)dx$ and the norm $||u|| \triangleq [\int_0^1 \rho(x)A(x)[u(x)]^2 dx]^{1/2}$ for all $u,v \in L_2(0,1)$ a new Hilbert space, denoted here by $\underset{\sim}{L}_2(0,1)$, is constructed.

Note that each of the sets D_a, D_b and D_c is an everywhere dense set in $\underset{\sim}{L}_2(0,1)$.

From now on all the considerations apply to $\underset{\sim}{L}_2(0,1)$.

Since each of the operators L_a, L_b and L_c, is a positive definite symmetric linear operator over its domain of definition (i.e., D_a, D_b or D_c), it follows that each of the operators L_a^{-1}, L_b^{-1} and L_c^{-1} exists and is a linear bounded self-adjoint operator which is defined over all of $\underset{\sim}{L}_2(0,1)$.

Now, since the system (1) with the initial conditions (5) has to obey one of the sets of boundary conditions (2), (3) or (4), the following equations have to be solved:

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = -L_a u + Qf(t) \\ u \in D_a \quad , \quad u(\cdot,0) = u_t(\cdot,0) = 0 \end{cases} \tag{13}$$

(follows from (1), (2) and (5)),

or

$$\begin{cases} \dfrac{\partial^2 v}{\partial t^2} = -L_b v + Qf(t) \\ v \in D_b \quad , \quad v(\cdot,0) = v_t(\cdot,0) = 0 \end{cases} \tag{14}$$

(follows from (1), (3) and (5)),

or

$$\begin{cases} \dfrac{\partial^2 w}{\partial t^2} = -L_c w + Qf(t) \\ w \in D_c \quad , \quad w(\cdot,0) = w_t(\cdot,0) = 0 \end{cases} \tag{15}$$

(follows from (1), (4) and (5)).

Note that the solution to each problem has been denoted by a different letter.

By operating with L_a^{-1} on both sides of (13), with L_b^{-1} on both sides of (14) and with L_c^{-1} on both sides of (15), the following equations are obtained:

$$\begin{cases} L_a^{-1} \dfrac{\partial^2 u}{\partial t^2} + u = L_a^{-1} Qf(t) \\ \\ u \in D_a \quad , \quad u(\cdot,0) = u_t(\cdot,0) = 0 \end{cases} \tag{16}$$

$$\begin{cases} L_b^{-1} \dfrac{\partial^2 v}{\partial t^2} + v = L_b^{-1} Qf(t) \\ \\ v \in D_b \quad , \quad v(\cdot,0) = v_t(\cdot,0) = 0 \end{cases} \tag{17}$$

$$\begin{cases} L_c^{-1} \dfrac{\partial^2 w}{\partial t^2} + w = L_c^{-1} Qf(t) \\ \\ w \in D_c \quad , \quad w(\cdot,0) = w_t(\cdot,0) = 0 \ . \end{cases} \tag{18}$$

Since (16) is equivalent to (13), (17) - to (14) and (18) - to (15), (16), (17) and (18) will be solved here.

The solutions to these equations are approximated by using the Method of Moments [4].

Let L^{-1} denote one of the operators L_a^{-1}, L_b^{-1} or L_c^{-1} according to the problem being solved.

At the first step the following sequence is defined:*

$$z_o = Q \quad z_{i+1} = L^{-1} z_i \quad , \quad i = 0,1,2,..., \tag{19}$$

which is given, in the case of (16), explicity by:

---

* It is assumed here, without loss of generality, that Q belongs to each of the sets D_a, D_b and D_c.

$$\left.\begin{aligned} z_{i+1}(x) = L_a^{-1} z_i = \int_0^x \frac{(x-\lambda)}{p(\lambda)} \int_0^\lambda (\lambda-\mu) q(\mu) z_i(\mu)\,d\mu\,d\lambda \; + \\ + B_1(z_i)\int_0^x (x-\lambda)\frac{\lambda}{p(\lambda)}\,d\lambda + B_2(z_i)\int_0^x \frac{(x-\lambda)}{p(\lambda)}\,d\lambda \\ 0 \le x \le 1 \end{aligned}\right\} \quad (20)$$

where $B_1(z_i)$ and $B_2(z_i)$ are the solutions of

$$\begin{pmatrix} \int_0^1 \frac{x\,dx}{p(x)} & \int_0^1 \frac{dx}{p(x)} \\ \int_0^1 \frac{(1-x)x}{p(x)}\,dx & \int_0^1 \frac{(1-x)}{p(x)}\,dx \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} -\int_0^1 \frac{1}{p(x)} \int_0^x (x-\lambda) q(\lambda) z_i(\lambda)\,d\lambda\,dx \\ -\int_0^1 \frac{(1-x)}{p(x)} \int_0^x (x-\lambda) q(\lambda) z_i(\lambda)\,d\lambda\,dx \end{pmatrix}, \quad (21)$$

in the case of (17) - by

$$\left\{\begin{aligned} z_{i+1}(x) = L_b^{-1} z_i = \int_0^x \frac{(x-\mu)}{p(\mu)} \int_0^\mu (\mu-\lambda) q(\lambda) z_i(\lambda)\,d\lambda\,d\mu \\ + \left\{x\int_0^1 \frac{(1-\mu)\mu}{p(\mu)}\,d\mu - \int_0^x \frac{(x-\mu)\mu}{p(\mu)}\,d\mu\right\} \int_0^1 (1-\lambda) q(\lambda) z_i(\lambda)\,d\lambda \\ - x\int_0^1 \frac{(1-\mu)}{p(\mu)} \int_0^\mu (\mu-\lambda) q(\lambda) z_i(\lambda)\,d\lambda\,d\mu \quad , \quad 0 \le x \le 1 \;, \end{aligned}\right. \quad (22)$$

and in the case of (18) - by

$$\left.\begin{aligned} z_{i+1}(x) = L_c^{-1} z_i = \int_0^x \frac{(x-\mu)}{p(\mu)} \int_0^\mu (\mu-\lambda) q(\lambda) z_i(\lambda)\,d\lambda\,d\mu \\ - \int_0^x \frac{(x-\mu)}{p(\mu)} \cdot \mu\,d\mu \int_0^1 q(\lambda) z_i(\lambda)\,d\lambda \; + \\ + \int_0^x \frac{(x-\mu)}{p(\mu)}\,d\mu \int_0^1 \lambda q(\lambda) z_i(\lambda)\,d\lambda \quad , \quad 0 \le x \le 1 \end{aligned}\right\} \quad (23)$$

In each of the cases

$$p(x) \triangleq E(x)I(x) \quad (24)$$

and

$$q(x) \triangleq \rho(x)A(x). \quad (25)$$

At the second step, for each of the systems (16), (17) or (18), the linear system

$$\begin{pmatrix} R_{oo} & R_{o1} & \cdots & R_{o,n-1} \\ R_{1o} & R_{11} & \cdots & R_{1,n-1} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ R_{n-1,o} & R_{n-1,1} & \cdots & R_{n-1,n-1} \end{pmatrix} \begin{pmatrix} \alpha_o \\ \alpha_1 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{n-1} \end{pmatrix} = - \begin{pmatrix} R_{on} \\ R_{1n} \\ \cdot \\ \cdot \\ \cdot \\ R_{n-1,n} \end{pmatrix} \tag{26}$$

where

$$R_{ij} \triangleq \int_o^1 q(x) z_i(x) z_j(x) dx \;, \tag{27}$$

is solved for $(\alpha_o, \alpha_1, \ldots, \alpha_{n-1})$.

At the third step, the following linear system, for a certain $f \in \Omega$ is solved:

$$\left\{ \begin{array}{l} \eta_o(t) - \alpha_o \dfrac{d^2 \eta_{n-1}(t)}{dt^2} = 0 \\[2ex] \dfrac{d^2 \eta_o(t)}{dt^2} + \eta_1(t) - \alpha_1 \dfrac{d^2 \eta_{n-1}(t)}{dt^2} = f(t) \\[2ex] \dfrac{d^2 \eta_1(t)}{dt^2} + \eta_2(t) - \alpha_2 \dfrac{d^2 \eta_{n-1}(t)}{dt^2} = 0 \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{d^2 \eta_{n-2}(t)}{dt^2} + \eta_{n-1}(t) - \alpha_{n-1} \dfrac{d^2 \eta_{n-1}(t)}{dt^2} = 0 \\[2ex] n_i(0) = 0 \quad \left. \dfrac{d\eta_i(t)}{dt} \right|_{t=0} = 0 \qquad i = 0,1,2,\ldots,n-1. \end{array} \right. \tag{28}$$

Denote by $\{z_i^a\}$ and $\{\eta_i^a\}$ the sequences generated by (20) and (26-28), respectively, by $\{z_i^b\}$ and $\{\eta_i^b\}$ the sequences generated by (22) and (26-28) respectively, and by $\{z_i^c\}$ and $\{\eta_i^c\}$ the sequences generated by (23) and (26-28) respectively.

The approximate solutions are represented respectively by [4].

$$u_n(x,t;f) = \sum_{k=o}^{n-1} \eta_k^a(t) z_k^a(x) \tag{29}$$

$$v_n(x,t;f) = \sum_{k=o}^{n-1} \eta_k^b(t) z_k^b(x) \tag{30}$$

$$w_n(x,t;f) = \sum_{k=o}^{n-1} \eta_k^c(t) z_k^c(x) \quad . \tag{31}$$

It is assumed that $z_o, z_1, \ldots, z_{n-1}$ are linearly independent functions for any approximation for each of the problems solved. If for some n, z_n is a linear combination of $z_o, z_1, \ldots, z_{n-1}$, then $u_n(x,t;f)$, $v_n(x,t;f)$ or $w_n(x,t;f)$ is an exact solution of the appropriate problem. For details see [4].

From [4] (pp. 19-21 and 103-104) it follows that

$$\lim_{n\to\infty} ||u_n(\cdot,t;f) - u(\cdot,t;f)|| = 0 \tag{32}$$

$$\lim_{n\to\infty} ||v_n(\cdot,t;f) - v(\cdot,t;f)|| = 0 \tag{33}$$

$$\lim_{n\to\infty} ||w_n(\cdot,t;f) - w(\cdot,t;f)|| = 0 \tag{34}$$

for all t ε [0,T] and for any f ε Ω.

For any f ε Ω, one can write:

$$\eta_k^a(t) = \int_o^t \phi_{kn}^a(t-\tau) f(\tau) d\tau \qquad k = 0,1,\ldots,n-1 \ , \tag{35}$$

$$\eta_k^b(t) = \int_o^t \phi_{kn}^b(t-\tau) f(\tau) d\tau \qquad k = 0,1,\ldots,n-1 \ , \tag{36}$$

$$\eta_k^c(t) = \int_o^t \phi_{kn}^c(t-\tau) f(\tau) d\tau \qquad k = 0,1,\ldots,n-1 \ . \tag{37}$$

Thus, for any f ε Ω, the approximate solutions are given by:

$$u_n(x,t;f) = \int_o^t \sum_{k=o}^{n-1} z_k^a(x)\phi_{kn}^a(t-\tau)f(\tau)d\tau \quad , \tag{38}$$

$$v_n(x,t;f) = \int_o^t \sum_{k=o}^{n-1} z_k^b(x)\phi_{kn}^b(t-\tau)f(\tau)d\tau \quad , \tag{39}$$

$$w_n(x,t;f) = \int_o^t \sum_{k=o}^{n-1} z_k^c(x)\phi_{kn}^c(t-\tau)f(\tau)d\tau \quad . \tag{40}$$

4. The Optimization Problem

It was shown in the previous section that u_n, (38), v_n, (39), and w_n, (40) are correspondingly the approximations to the solutions of the three problems defined above. It is clear that the same procedure of optimization can be applied to these three cases. For this reason the general designation of the solutions, u, will be used in the following treatment.

Denote by f^*, $f^* \varepsilon \Omega$, a control for which $I(f^*) \leq I(f)$ for any $f \varepsilon \Omega$. In this section the problem of the computation of f^* is discussed.

In the following, the discussion will be restricted to functions $u_d(x)$ which can be written in the form

$$u_d(x) = \sum_{k=o}^{\infty} \gamma_k z_k(x). \tag{41}$$

This restriction follows from the fact that, for any $f \varepsilon \Omega$, the solution, u, belongs to the sub-space of $\underset{\sim}{L}_2(0,1)$ which is spanned by $z_o, z_1, \ldots, z_n, \ldots$, as has been shown in the previous section.

Now, the functional $I_n(f)$ is defined as

$$I_n(f) \triangleq \int_o^1 [\sum_{k=o}^{n-1} (\gamma_k - \int_o^T \phi_{kn}(T-t)f(t)dt)z_k(x)]^2 dx \quad , \tag{42}$$

and is rewritten in the form

$$I_n(f) = I_n(0) - 2\int_o^T F_n(t)f(t)dt + \int_o^T \int_o^T R_n(t,s)f(t)f(s)dsdt \tag{43}$$

where
$$F_n(t) \triangleq \sum_{i,j=o}^{n-1} \gamma_i R_{ij} \phi_{jn}(T-t), \tag{44}$$

and
$$R_n(t,s) \triangleq \sum_{i,j=o}^{n-1} \phi_{in}(T-t) R_{ij} \phi_{jn}(T-s) \quad . \tag{45}$$

Denote by f_n^*, $f_n^* \varepsilon \Omega$, a control for which $I_n(f_n^*) \leq I_n(f)$ for any $f \varepsilon \Omega$. It can be shown [5] that

$$\lim_{n\to\infty} I_n(f_n^*) = I(f^*) \ . \tag{46}$$

Assume that, for n large enough, a control $f_n^* \in \Omega$ is known. Then, using the scheme described in Section 2, the function $u_n(x,t;f_n^*)$ is given. The substitution of $u_n(x,t;f)$ for any $f \in \Omega$, in one of the equations (16), (17) or (18) which corresponds to the problem being solved yields an error which will be denoted here by $\varepsilon_n(x,t;f)$, i.e.

$$L^{-1} \frac{\partial^2 u_n}{\partial t^2} + u_n - L^{-1}Qf(t) \overset{\Delta}{=} \varepsilon_n(x,t;f) \ . \tag{47}$$

It can be shown, in the same manner as in [6], that $\varepsilon_n(x,t;f)$ is given by

$$\varepsilon_n(x,t;f) = \frac{d^2\eta_{n-1}(t)}{dt^2} (z_n(x) + \alpha_o z_o(x) + \alpha_1 z_1(x) + \ldots + \alpha_{n-1} z_{n-1}(x)). \tag{48}$$

By computing $\varepsilon_n(x,t;f_n^*)$ the validity of the solution $u_n(x,t;f_n^*)$ can be checked.

REFERENCES

[1] V. Komkov, "The Optimal Control of a Transverse Vibration of a Beam", J. SIAM Control, Vol. 6, No. 3, pp. 401-421, 1968.

[2] A. V. Balakrishnan, "Optimal Control Problems in Banach Spaces", J. SIAM Control, Vol. 3, No. 1, 1965.

[3] S. G. Mikhlin, "The Problem of the Minimum of a Quadratic Functional", Holden-Day, Inc., San Francisco, pp. 42-45, 1965.

[4] Yu. V. Vorobyev, "Method of Moments in Applied Mathematics", Gordon and Breach Science Publishers, New York, 1965.

[5] Y. Yavin and Y. Rasis, "Optimal Control for Two Classes of Distributed Parameter Systems", Presented at the Second Annual Houston Conference on Circuits, Systems, and Computers, April 20-21, 1970.

[6] Y. Yavin and Y. Rasis, "The Bounded Energy Optimal Control for a Class of Heat Conduction Systems", Int. J. Control, Vol. 11, No. 1, pp. 153-164, January 1970.

[7] V. F. Demyanov and A. M. Rubinov, "The Minimization of a Smooth Convex Functional on a Convex Set", J. SIAM Control, Vol. 5, No. 2, pp. 280-294, 1967.

A DISTRIBUTED PARAMETER NETWORK APPROACH TO THE ANALYSIS OF ESTUARY POLLUTION

B. P. BUTZ AND R. FISCHL
Department of Electrical Engineering
Drexel University
Philadelphia, Pennsylvania 19104

ABSTRACT

Distributed parameter network theory is used to find the pollution concentration along an estuary. An exact functional form of the pollution concentration is obtained for each segment of the estuary by modeling each segment as an exponentially tapered distributed parameter network. Thus the estuary is modeled as a cascade of distributed parameter sections, each of which can be represented by a two-port device with known transmission parameters. The frequency and transient responses of the estuary can be obtained readily on a digital computer. A numerical example is given.

INTRODUCTION

The simulation of water pollution dynamics is based on a generalized diffusion equation which describes the rate of change of concentration, C, of dissolved pollutant, namely[1-3]

$$\frac{\partial C}{\partial t} = \underline{\nabla} \cdot E(\underline{\nabla} C) - \underline{\nabla} \cdot \underline{\mu} C + R \tag{1}$$

The first term on the right side of Eq. (1) is the diffusion term and E is the diffusion coefficient. The second term represents advection; the coefficient μ is the velocity. The last term represents the reactive effect which is considered to be a first order expression

$$R = \pm mC \pm \Sigma S$$

where m is the rate of decay or regeneration of the pollutant and ΣS is the sum of all sources or sinks. Note that all the coefficients vary spatially and temporally.

In general the variations in the vertical and lateral direction are considered negligible because of the confining boundaries of the waterway. Hence the diffusion equation becomes

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left[E(x,t)\frac{\partial C}{\partial x}\right] - \frac{\partial \mu C}{\partial x} \pm mC \pm \Sigma S \tag{2}$$

In terms of the estuary parameters this equation becomes

$$\frac{\partial C}{\partial t} = \frac{1}{A(x,t)}\left\{\frac{\partial}{\partial x}\left[K(x,t)\,A(x,t)\,\frac{\partial C}{\partial t}\right] - \frac{\partial}{\partial x}\left[Q(x,t)C\right]\right\} \pm m(x,t)C \pm \Sigma S(x,t) \tag{3}$$

where C = concentration level (mg/L)

A = cross-sectional area (square ft.)

Q = net fresh water inflow (cu. ft/sec)

m = rate of decay or regeneration of concentration (1/day)

K = eddy diffusion coefficient (mi^2/day)

t = time (days)

x = distance along estuary (ft)

S = all sources and sinks of concentration independent of the concentration (mg/L-day)

Equation (3) has the general notion of pollution "concentration level" as the dependent variable. The primary indicator of water quality is taken to be the amount of dissolved oxygen (DO). The maximum value of dissolved oxygen, called the saturation level, is a function of the temperature and chlorine concentration of the estuary. When oxygen is furnished to waste material containing bacteria, decomposition occurs until the oxygen demand is satisfied. The amount of oxygen absorbed during this process is known as the biochemical oxygen demand (BOD). Equation (3) represent either the dissolved oxygen (DO) or the biochemical oxygen demand (BOD) by properly defining the variables m and S.

To obtain the BOD concentration equation set

$m = -d$, the decay rate of BOD (day^{-1})

$S = J$, the amount of effluent BOD discharged into the estuary (mg/L-day)

and $C = L$, the BOD concentration (mg/L)

Then for BOD concentration Eq. (3) becomes,

$$\frac{\partial L}{\partial t} = \frac{1}{A(x,t)} \left\{ \frac{\partial}{\partial x} \left[K(x,t)\ A(x,t)\ \frac{\partial L}{\partial t} \right] - \frac{\partial}{\partial x} \left[Q(x,t)\ L \right] \right\} - d(x,t)\ L \pm \sum J(x,t) \tag{4}$$

To obtain the DO concentration equation set

$m = -r$, the rearation rate from the atmosphere (day^{-1})

$H =$ the DO concentration (mg/L)

and $S = -d\ L + P - H_s$

where P = all independent sources and sinks of DO (mg/L-day)

H_s = satuaration constant of DO (mg/L)

d, L are previously defined

Then for DO concentration Eq. (3) becomes,

$$\frac{\partial H}{\partial t} = \frac{1}{A(x,t)} \left\{ \frac{\partial}{\partial x} \left[K(x,t)\ A(x,t)\ \frac{\partial}{\partial x} \left[Q(x,t)\ H \right] \right\} \tag{5}$$

$$+ r(x,t)\ [H_s - H] - d(x,t)\ L \mp \sum P(x,t) \tag{5}$$

The determination of the distribution of pollution in an estuary as a function of position and time involves the solution of the pair of second order partial differential equations given by Eq. 4 and Eq. 5. It should be noted that this pair of equations are coupled in a sense that one of the sources of the DO equation is the BOD.

In this paper we are primary concerned with the solution of Eq. 3. Since the boundaries and coefficients of Eq. 3 are poorly defined, only solutions to the average concentration are sought. Even this problem is a difficult one, so that various researches sought solutions under different assumptions, e.g. the absence of the reaction term.[4] One of the more successful methods of solving Eq. 3 is by using a finite-difference approximation,[5] i.e. the lumped-parameter model of the estuary. An approximate solution of this model has been obtained by using the ECAP computer program[6]. This paper considers an alternate model which is based on distributed parameter networks.

DISTRIBUTED PARAMETER MODEL OF THE ESTUARY

We begin the development of the distributed parameter model of the estuary with the equations of a nonuniform transmission line with a distributed independent current source, namely:

$$\frac{\partial v(x,t)}{\partial x} = -r(x)\ i(x,t) \tag{6}$$

$$\frac{\partial i(x,t)}{\partial x} = -[g(x)v(x,t) + C(x)\ \frac{\partial v(x,t)}{\partial t}\ ] + F(x,t) \tag{7}$$

where
v(x,t) = the voltage along the distributed network (volts)
i(x,t) = the current in the distributed network (amps)
r(x) = resistance per unit length (ohms/meter)
c(x) = capacitance per unit length (farads/meter)
g(x) = conductance per unit length (mhos/meter)
x = distance along the line (meters)
t = time (seconds)
f(x,t) = distributed independent current source (amps/meter)

Writing Eq. (6) and Eq. (7) in terms of the voltage, yields

$$\frac{\partial^2 v}{\partial x^2} = (\frac{1}{r}\frac{\partial r}{\partial x})\ \frac{\partial v}{\partial x} + (rg)\ v + rc\ \frac{\partial v}{\partial t} - rF \tag{8}$$

Expanding Eq. (3) and rearranging the terms, gives

$$\frac{\partial^2 C}{\partial x^2} = -\left(\frac{1}{KA}\frac{\partial KA}{\partial x} - \frac{Q}{KA}\right)\frac{\partial C}{\partial x} + \frac{1}{KA}\left(\frac{\partial Q}{\partial x} - mA\right)C + \frac{1}{K}\frac{\partial C}{\partial t} - \frac{1}{K}\sum S \tag{9}$$

Comparison of Eq. (8) with Eq. (9) shows that if the estuary variables A, Q, S, K and m are averaged over a tidal cycle (which is usually the case), the network parameters can be related to the estuary parameters as shown in Table I. Once the analogy between the estuary and the $\overline{\text{RCG}}$ network[*] has been established, the techniques of $\overline{\text{RCG}}$ network analysis can be applied to the analysis of the pollution distribution of the estuary.

The exact analysis of the $\overline{\text{RCG}}$ network depends on the type of non-uniform taper specified. This implies that one must know the exact functional expressions of the estuarine parameters as a function of position. In general these expressions cannot be determined. One can circumvent this problem by dividing the estuary into a finite number of segments, each of which contains spacially constant estuary parameters A, K, Q and m and where all independent sources and sinks are applied at the boundary between segments.

Let us assume that the parameters K, A, Q and m are constant over each estuary segment, then from Table I, the network parameters of Eqs. (6) and (7) are given by

$$\begin{aligned} r(x) &= r_o e^{\alpha x} \\ g(x) &= g_o e^{-\alpha x} \\ c(x) &= c_o e^{-\alpha x} \end{aligned} \tag{10}$$

$$\text{where} \quad \alpha = \frac{Q}{KA} \qquad c_o = A$$
$$r_o = \frac{1}{KA} \qquad g_o = -mA \tag{11}$$

Hence each estuary segment can be modeled by an exponentially tapered $\overline{\text{(ERCG)}}$ network. Furthermore, the external characteristic of each estuary segment can be considered to be identical to the input-output characteristics of a two-port $\overline{\text{ERCG}}$ network. Thus, the kth segment of the estuary can be simulated by a two-port, shown in Figure 1, whose transmission characteristics are given by[**]

$$\begin{bmatrix} C_{k1} \\ I_{k1} \end{bmatrix} = \begin{bmatrix} T_k \end{bmatrix} \begin{bmatrix} C_{k2} \\ -I_{k2} \end{bmatrix} \tag{12}$$

---

[*] The notation $\overline{\text{RCG}}$ network is used for a tapered distributed network whose equivalent circuit can be represented by a series resistance branch and parallel capacitance-conductance branches. The exponentially tapered $\overline{\text{RCG}}$ network will be denoted by $\overline{\text{ERCG}}$ network.

[**] In the following discussion it is assumed that all independent sources and sinks are applied at the boundary between segments.

where C_{k1} and C_{k2} are the concentrations at the input and output of the kth segment, respectively. The variables I_{k1} and I_{k2} are related to the spacial derivative of the concentration.

The transmission matrix, $[T_k]$, is given by [7]

$$[T_k] = \left[\begin{array}{c|c} e^{-\alpha D/2}(\cosh\phi + \frac{\alpha D}{2\phi}\sinh\phi) & e^{\alpha D/2}\ \frac{D\sinh\phi}{KA\ \phi} \\ \hline e^{-\alpha D/2}\ \frac{KA}{D\phi}\ \phi^2 - \frac{\alpha^2 D^2}{4}\ \sinh\phi & e^{\alpha D/2}(\cosh\phi - \frac{\alpha D}{2\phi}\sinh\phi) \end{array}\right] \tag{13}$$

where

$$\begin{aligned} \phi &= \sqrt{D^2\left(\frac{\alpha^2}{4} + \frac{s \mp m}{K}\right)} \\ \alpha &= \frac{Q}{KA} \\ D &= \text{length of segment (feet)} \\ s &= \text{complex frequency variable} \end{aligned} \tag{14}$$

For the special case of negligible sources and sinks along the estuary the concentration at the output terminals of any segment of the estuary can be obtained in terms of the first segment by viewing the estuary as a cascade connection of the two-port networks, namely

$$\begin{bmatrix} C_{11} \\ I_{11} \end{bmatrix} = \begin{bmatrix} T_n \end{bmatrix} \begin{bmatrix} T_{n-1} \end{bmatrix} \cdots \begin{bmatrix} T_2 \end{bmatrix} \begin{bmatrix} T_1 \end{bmatrix} \begin{bmatrix} C_{n2} \\ -I_{n2} \end{bmatrix} \tag{15}$$

where C_{11} and I_{11} are the values at the input boundary of the first segment C_{n2} and I_{n2} are the values at the output of the nth segment; and $[T_k]$ is the transmission matrix of Eq. (13) for the kth segment.

To obtain the concentration within a particular segment (e.g., the nth segment) it is required to know the values of C_{n1} and I_{n1} at the input boundary. Then it can be shown that any point within the nth segment, the concentration is given by

$$C_n(x) = e^{\alpha x/2}\ [C_{n1}\cosh\left(\frac{\phi x}{D}\right) - \left(\frac{2I_{n1} + QC_{n1}}{2KA}\right)\sinh\left(\frac{\phi x}{D}\right)] \tag{16}$$

It should be noted that the transmission matrix, T_k, defined by Eq. (13) gives the transfer ratio of the pollution concentration for the kth estuary segment in functional form. Hence, the poles and zeros of the transfer ratio are known, and the frequency and transient responses at any point along the estuary can be obtained.

NUMERICAL EXAMPLE

This section illustrates the application of the distributed parameter model in the analysis of the pollution concentration in an estuary. An ideal situation is simulated so as to obtain the transfer characteristics of an estuary segment which is modeled by an $\overline{\mathrm{ERCG}}$ network. The concentration transfer ratio under the assumption that the output rate of change of concentration is zero (i.e. open circuit voltage gain) is given by

$$T(s) = \frac{C_{k2}}{C_{k1}}(s) = \frac{1}{t_{11}}(s) = \frac{e^{\alpha D/2}}{\cosh\phi + \frac{\alpha D}{2\phi}\sinh\phi} \tag{17}$$

where ϕ is defined by Eq. (14). A frequency response of the estuary segment is shown in Figure 2. The estuarine parameters used are given in Table II. Recall that the transfer function, T(s), of Eq. (17) can be written as[7]

$$T(s) = \frac{T_o}{\prod_{i=1}^{\infty}\left(1+\frac{s}{P_i}\right)} \tag{18}$$

where T_o is the zero frequency value of $T(j\omega)$, and P_i is the ith pole of T(s). The first six poles of this transfer functions have been evaluated and are given in Table II.

To obtain the step response shown in Figure 3, the equivalent dominant pole approximation[7] was used. The transfer function T(s) was approximated by

$$T_e(s) = \frac{T_o}{1+\frac{s}{P_e}}\, e^{-\frac{\theta_e}{P_e}s} \tag{19}$$

where P_e is the equivalent dominant pole which represents the 3db frequency point of $|T(j\omega)|$; and θ_e is the excess phase which represent the phase (in radians) in excess of $\frac{\Pi}{4}$ radians at P_e. The calculated values are $P_e = 16.704$ rad/day and $\theta_e = -0.953$ radians. The unit step response based on the equivalent dominant pole approximation, $T_e(s)$, is given by

$$C_{2k}(t) = T_o\left[1 - e^{-\left(t - \frac{\theta_e}{P_e}\right)}\right] \tag{20}$$

and is plotted in Figure 3. It should be noted that this response represents the first approximation to the step response (see Fig. 2), and can be improved by assuming a higher order transfer function approximation. Furthermore the ideal case considered herein should be modified to include the effect of terminating the estuary segment by its characteristic impedance.

CONCLUSIONS AND FUTURE WORK

In this paper we have shown that the pollution concentration along an estuary is analogous to the voltage distribution along an $\overline{\mathrm{ERCG}}$ network. An exact functional form of the pollution concentration ratio has been obtained for each segment of the estuary. We have obtained the transient response for one section of the estuary. It is felt that the distributed parameter model will yield more accurate results than those obtained previously[5,6], because fewer constraints have been placed on this model. In the immediate future, it is planned to obtain the complete response of the estuary with independent sources included.

TABLE I:
RELATION BETWEEN ESTUARY AND NETWORK PARAMETERS

Network Parameters	Estuary Parameters
r (ohms/meter)	$\frac{1}{KA} \exp [\int \frac{Q}{KA} dx]$ (time/length^4)
c (farads/meter)	$A \exp [- \int \frac{Q}{KA} dx]$ (length^2)
g (mhos, meter)	$(\frac{\partial Q}{\partial x} - mA) \exp [- \int \frac{Q}{KA} dx]$ (length^2/time)
v (volts)	C (concentration)
F (amps/meter)	$A \exp [- \int \frac{Q}{KA} dx] \Sigma S$ $(\frac{\text{concen.-length}^2}{\text{time}})$

TABLE II:
SUMMARY OF NUMERICAL EXAMPLE

A) Estuary Parameters

Parameter	Value	Units
Cross Sectional Area, A	6500	sq. ft.
Net Fresh Water Inflow, Q	8284	cu. ft./sec.
Segment Length, D	21000	feet
Eddy Diffusion Coefficient, K	4	sq. mi./day
Decay Coefficient, m	.1023	day^{-1}

B) Poles of Transfer Function T(s)

$S_1 = -30.114$

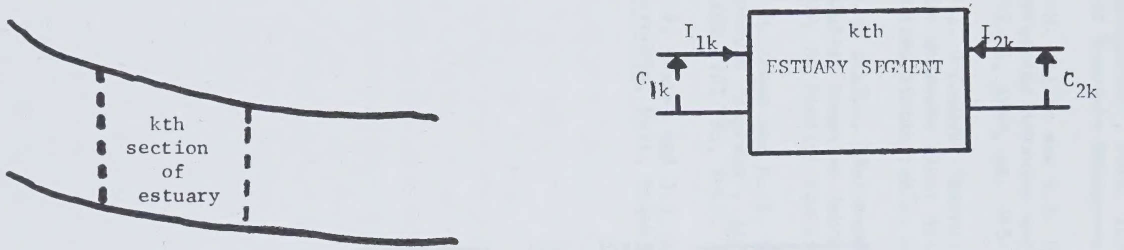
$S_2 = -39.008$

$S_3 = -53.399$

$S_4 = -73.039$

$S_5 = -97.802$

$S_6 = 127.629$


kth
section
of
estuary
I_{1k}
C_{1k}
kth
ESTUARY SEGMENT
I_{2k}
C_{2k}

a) Segmentation of Estuary

b) The Estuary Simulated as a Two-Port Device

FIGURE 1

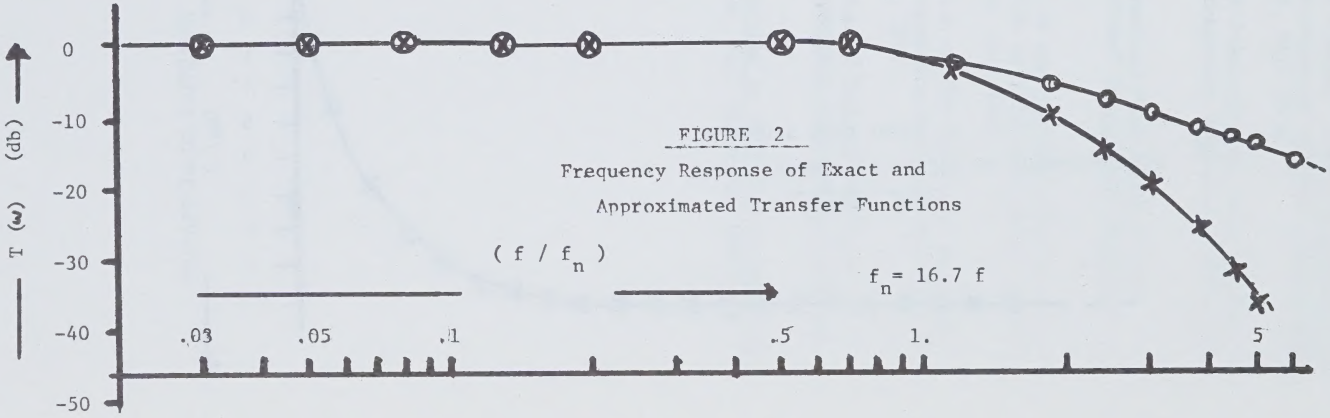


FIGURE 2

Frequency Response of Exact and Approximated Transfer Functions

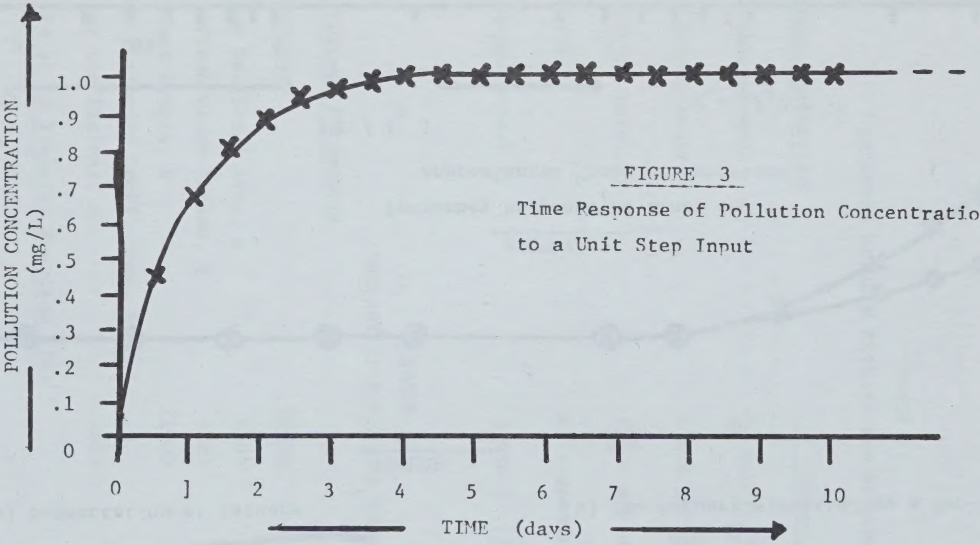


FIGURE 3

Time Response of Pollution Concentration to a Unit Step Input

REFERENCES

[1] D.J. O'Connor, L.L. Falk and R.G.E. Franks, "Analog Computation of the Dispersion of Pollutants in the Delaware Estuary", Proc. 18th Ind. Waste Conf., Purdue Univ., Ext. Sec. 1964.

[2] R.V. Thomann, "Systems Analysis and Simulation in Water Quality Management", Proc. IBM Scientific Computing Symposium, Water and Air Resource Management, 1967.

[3] J.M. Jeglic and G.D. Pence, "Mathematical Simulation of the Estuarine Behavior and its Applications" Socio-Econ. Plan Sci., Vol. 1, 1968, pp. 363-389.

[4] D.J. O'Connor, "Report on Analysis of the Dye Diffusion Data in the Delaware River Estuary-Evaluation of Diffusion Coefficients", United States Public Health Service, Aug. 1962.

[5] J.M. Jeglic, "Mathematical Simulation of the Estuarine Behavior" Digital Computer Technology and Programming Analysis Memo No. 1032, Rev. A, General Electric Co., July, 1967.

[6] J.M. Leeds and H. H. Bybee, "Solution of Estuary Problems and Network Programs", Journal of the Sanitary Engineering Division, ASCE, Vol. 93, 1967.

[7] M.S. Ghausi and J.J. Kelly, Introduction to Distributed Parameter Networks, Holt, Rinehart and Winston, Inc. New York, 1968.

ON THE DESIGN OF A FOUR-PORT DIRECTIONAL FILTER

MASAO HIBINO
Transmission Division, Nippon Electric Co.,Ltd.
Kawasaki, Japan

ABSTRACT

In this paper, a theory which makes it possible to design a directional filter without use of compensating networks is presented.
By this method, the same electrical performance can be achieved with less number of elements than by conventional method. A directional filter more reliable and smaller in size can be obtained.

I. Introduction

Two-way single amplifier repeaters which were proposed by Darlington are widely used in transmission systems such as submarine cable systems. Desoer [1]has presented a theory and a design method of a four-port directional filter used in this type of repeaters. His method is based on the theory of four-port constant R network, but when it is applied to actual design, it becomes necessary to connect a one-port network at each of the four-ports for the following two reasons.
One is to compensate the susceptance which appears because of approximations in the design procedure. The other reason is to ensure enough loss between two ports connected to the amplifier. Decrease of loss makes the repeater unstable because the directional filter constitutes a feedback path of the amplifier. As was pointed in Desoer's paper, the loss can be made infinite in theory by inserting an ideal phase inversion transformer. In practice, however, the loss becomes considerably small because of inaccuracy of the element values and existence of parasitics.

In this paper, we propose a theory which makes it possible to design a directional filter without use of compensating networks. By this method, the same electrical performance can be achieved with less number of elements than by Desoer's method. Consequently, we can have filters more reliable and smaller in size.

II. Analysis of the Four-Port Network

We consider the configuration of Fig.- 1.
It is assumed that
1) The two low-pass filters are identical.
2) The two high-pass filters are identical.

3) The four filters are symmetric.

4) The four filters are reactive.

Let the admittance matrices of the low-pass and the high-pass filters be, respectively,

$$\begin{bmatrix} y_\ell & y_{\ell t} \\ y_{\ell t} & y_\ell \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} y_h & y_{ht} \\ y_{ht} & y_h \end{bmatrix}$$

Then, the scattering matrix of the four-port is

$$S = \begin{bmatrix} \rho & 0 & \tau_\ell & \tau_h \\ 0 & \rho & -\tau_h & \tau_\ell \\ \tau_\ell & -\tau_h & \rho & 0 \\ \tau_h & \tau_\ell & 0 & \rho \end{bmatrix} \tag{1}$$

$$\begin{aligned} \text{where} \quad \rho &= S_{11} = \frac{1 - y_o}{1 + y_o} \\ \tau_\ell &= S_{13} = \frac{2}{1+y_o}\left(\frac{-y_{\ell t}}{1+y}\right) \\ \tau_h &= S_{14} = \frac{2}{1+y_o}\left(\frac{-y_{ht}}{1+y}\right) \\ y_o &= y - \frac{y_{ht}^2 + y_{\ell t}^2}{1+y} \end{aligned} \tag{2}$$

and

$$y = y_h + y_\ell$$

y_o expresses the input admittance at each port.

Since the network is reactive,

$$\begin{aligned} S_{13}\widetilde{S_{13}} + S_{14}\widetilde{S_{14}} &= 1 - S_{11}\widetilde{S_{11}} \\ &= \left(\frac{2}{1+y_o}\right)\left(\frac{2}{1+\widetilde{y_o}}\right) \mathrm{Ev}\, y_o \end{aligned} \tag{3}$$

where $\sim$ denotes the substitution of p by -p, and Ev y_o expresses even part of y_o.

Let G(p) be even part of y_o, then reminding that y_h and y_ℓ are odd rational functions of p, we have

$$G(p) = \mathrm{Ev}\, y_o = \mathrm{Ev}\, \frac{1-\rho}{1+\rho} = \frac{y_{ht}^2 + y_{\ell t}^2}{y^2 - 1} \tag{4}$$

or

$$y_{ht}^2 + y_{\ell t}^2 = (y^2 - 1)\, G(p) \tag{5}$$

The above relations (4), (5) are important for our problem.

The example is designed so as to have the same electrical performances

Let Ω_H, Ω_L and Ω_T denote the upper pass-band, the lower pass-band and the transition-band respectively.

Combining (2) and (5), we have

$$\begin{aligned} \frac{1}{|S_{14}|^2} &= \left(\frac{1+y_o}{2}\right)\left(\frac{1+\tilde{y}_o}{2}\right) \cdot \frac{y_{ht}^2 + y_{\ell t}^2}{y_{ht}^2\, G(p)} \\ \frac{1}{|S_{13}|^2} &= \left(\frac{1+y_o}{2}\right)\left(\frac{1+\tilde{y}_o}{2}\right) \cdot \frac{y_{ht}^2 + y_{\ell t}^2}{y_{\ell t}^2\, G(p)} \end{aligned} \qquad (6)$$

Since G(p) is close to unity in the pass-bands, (6) has the following approximations:

$$\begin{aligned} \frac{1}{|S_{14}|^2} &\simeq 1 + \frac{y_{\ell t}^2}{y_{ht}^2} \qquad \text{for } \omega \in \Omega_L \\ \frac{1}{|S_{13}|^2} &\simeq 1 + \frac{y_{ht}^2}{y_{\ell t}^2} \qquad \text{for } \omega \in \Omega_H \end{aligned} \qquad (7)$$

By the same discussions with the reference [1],

$$\begin{aligned} |y_{\ell t}|^2 &\gg |y_{ht}|^2 \qquad \text{for } \omega \in \Omega_L \\ |y_{ht}|^2 &\gg |y_{\ell t}|^2 \qquad \text{for } \omega \in \Omega_H \end{aligned} \qquad (8)$$

and

$$\begin{aligned} y_{\ell t}^2 + y_{ht}^2 &\simeq y_{\ell t}^2 \qquad \text{for } \omega \in \Omega_L \\ y_{\ell t}^2 + y_{ht}^2 &\simeq y_{ht}^2 \qquad \text{for } \omega \in \Omega_H \end{aligned} \qquad (9)$$

Putting $\varphi_h = -\dfrac{1}{y_{ht}}$ and $\varphi_\ell = -\dfrac{1}{y_{\ell t}}$,

we have

$$\begin{aligned} \frac{1}{|S_{14}|^2} &\geqq \frac{1}{|y_{ht}^2|} = |\varphi_h|^2 \qquad \text{for } \omega \in \Omega_L \\ \frac{1}{|S_{13}|^2} &\geqq \frac{1}{|y_{\ell t}^2|} = |\varphi_\ell|^2 \qquad \text{for } \omega \in \Omega_H \end{aligned} \qquad (10)$$

Let us consider the requirements imposed on the above functions S_{11} (G(p)), S_{13}, S_{14} and S_{34}, etc.

For the filters, it is necessary that the loop attenuation between ports 3 and 4 is enough large to decrease the pass-band ripple of a repeater. And also, in the transition-band, the filters, at least, must have larger attenuation than amplifier's gain.

Now, the sensitivity of the operating transfer function between the ports 3 and 4 with respect to the terminating resister r_1 at the port 1 is

$$\left|\frac{\partial S_{34}}{\partial r_1}\right|^2 = \frac{1}{4}\left|S_{13}\right|^2 \cdot \left|S_{14}\right|^2 \quad \text{on the } j\omega \text{ axis.}$$

In order to ensure the stability of a repeater, the sensitivity is required to be small.

Since the high return loss is required in Ω_H and Ω_L, $\left|S_{11}\right|^2 \simeq 0$ for $\omega \in \Omega_H \cup \Omega_L$. In Ω_H, $\left|S_{13}\right|^2$ is very small in comparison with a magnitude of $\left|S_{14}\right|^2$ and, also, in Ω_L, $\left|S_{14}\right|^2$ is very small in comparison with a magnitude of $\left|S_{13}\right|^2$. Accordingly, the sensitivity which is proportional to the product of $\left|S_{13}\right|^2$ and $\left|S_{14}\right|^2$, becomes small.

In Ω_T, however, the magnitude of $\left|S_{13}\right|^2$ and $\left|S_{14}\right|^2$ are comparable, hence, in order to keep the sensitivity small, the sum $\left|S_{13}\right|^2 + \left|S_{14}\right|^2$ must be small.

Then, the requirements for the filters are summarized as follows;

1) $G(p)$ must be close to unity so that the filters have high return loss in the pass-bands. (equation (4))

2) S_{13} and S_{14} (S_{23} and S_{24}) must be so determined that the filters have small sensitivity of the operating transfer function S_{34} between the ports 3 and 4.
   This is reflected to $G(p)$ by the relation(4).

3) S_{13} and S_{14} (S_{23} and S_{24}), also, must be so determined that the filters have enough attenuation in all bands.

III Network Realization

In (5), we can find that the sum $y_{ht}^2 + y_{\ell t}^2$ must have maxima of value $-G(p)$ on the $j\omega$ axis.

From (9), y_{ht}^2, for $\omega \in \Omega_H$, has maxima of value $-G(p)$, which means that $|\varphi_h|^2$ oscillates between 0 and $\frac{1}{G(p)}$. The same for $y_{\ell t}^2$ holds for $\omega \in \Omega_L$.

Desoer's theorem for the necessary and sufficient conditions of physical realizability is modified and stated as follows:

1) $y_{\ell t}$ and y_{ht} must be odd rational functions of p, having simple poles located on the $j\omega$ axis.

2) The $j\omega$ poles of $y_{\ell t}$ are distinct from those of y_{ht}.

3) For $p = j\bar{\omega}$,

$$y_{\ell t}^2 + y_{ht}^2$$

   must have maxima of value $-G(p)$.

The condition 3) is different from Desoer's theorem and is important for our problem. Under these conditions, we consider the physical realization.

The example is designed so as to have the same electrical performances

Let the partial fraction expansions of $y_{\ell t}$ and y_{ht} be, respectively,

$$y_{\ell t} = \sum_{-n}^{+n}{}' \frac{k_i}{P - j\omega_i} + \frac{k_o}{p}$$

$$y_{ht} = \sum_{-m}^{+m}{}' \frac{h_k}{P - jq_k} + h_o p \tag{11}$$

where the prime indicates that the terms for $n, m = o$ are omitted

$\omega_i = -\omega_{-i}$ $\quad (i = 1, 2, \ldots, n)$

$q_k = -q_{-k}$ $\quad (k = 1, 2, \ldots, m)$

$\omega_i \neq q_k$ $\quad (i = 1, 2, \ldots, n; \; k = 1, 2, \ldots, m)$

Now, let us consider the relation (5),

$$y^2_{\ell t} + y^2_{ht} = (y^2 - 1)\, G(p) \tag{12}$$

When (12) is expanded in Laurent series around the point $p = j\omega_i$, $p = jq_k$, $p = 0$ and $p = \infty$, the dominant term of the left hand side must match the dominant term of the right hand side.

Therefore, y can be written as:

$$y = \sum_{-n}^{+n}{}' \frac{k_i{}'}{p - j\omega_i} + \sum_{-m}^{+m}{}' \frac{h_k{}'}{p - jq_k} + \frac{k_o{}'}{p} + h'_o\, p \tag{13}$$

Then, the residue matching conditions are as follows:

$$\begin{aligned} k'^2_i \; G(p) &= k_i^2 \text{ at } p = j\omega_i, \quad k'^2_o \; G(p) = k_o^2 \text{ at } p = 0 \\ h'^2_k \; G(p) &= h_k^2 \text{ at } p = jq_k, \quad h'^2_o \; G(p) = h_o^2 \text{ at } p = \infty \end{aligned} \tag{14}$$

The following conditions must be added so that the low-pass and high-pass filters are reactive.

$$\begin{aligned} k'^2_i - k_i^2 &\geq 0 \qquad k'^2_o - k_c^2 \geq 0 \\ h'^2_k - h_k^2 &\geq 0 \qquad h'^2_o - h_o^2 \geq 0 \end{aligned} \tag{15}$$

Consequently, we have

$$\begin{aligned} k'^2_i (1 - G(p)) &\geq 0 \text{ at } p = j\omega_i, \; k'^2_o (1 - G(p)) \geq 0 \text{ at } p = 0 \\ h'^2_k (1 - G(p)) &\geq 0 \text{ at } p = jq_k, \; h'^2_o (1 - G(p)) \geq 0 \text{ at } p = \infty \end{aligned} \tag{16}$$

Combining above conditions, we have

$G(p) \leq 1$ at $p = j\omega_i, p = jq_k, p = 0$ and $p = \infty$.

$G(p) = 1$ at $p = j\omega_i$, $p = jq_k, p = 0$ and $p = \infty$, satisfies the minimum susceptance condition. This requirement means that the zeros of φ_ℓ, φ_h must exist at the point of $G(p) = 1$.

So, we can have the reactance functions y_ℓ and y_h by

$$y_\ell = \sum_{-n}^{+n}{}' \frac{|k_i|}{p-j\omega_i} + \frac{|k_o|}{p}$$

$$y_h = \sum_{-m}^{+m}{}' \frac{|h_k|}{p-jq_k} + |h_o|\, p \tag{17}$$

IV. Design Procedure

The actual design procedure together with an example is shown as follows:

1) Determine $G(p)$ so that it satisfies the requirements of the return loss in the pass-bands, at the same time, gives the small sensitivity in the transition-band.
   In order to determine $G(p)$, we consider reference filters[2] which satisfy the above requirements.

   Let ρ_H and ρ_L be the reflection coefficients of the reference high-pass and low-pass filters, respectively, and obtain ρ as follows:

   $$\rho = \rho_H \cdot \rho_L \tag{18}$$

   where $\rho = \dfrac{1 - y_o}{1 + y_o}$, $\rho_H = \dfrac{1 - Y_H}{1 + Y_H}$ and $\rho_L = \dfrac{1 - Y_L}{1 + Y_L}$.

   Y_H, Y_L express the admittances of the reference high-pass and low-pass filters.

   From the relation $G(p) = \mathrm{Ev}\ y_o$, we have

   $$G(p) = \mathrm{Ev}\ \frac{Y_H + Y_L}{1 + Y_H Y_L} \tag{19}$$

   Fig.- 2 shows a given $G(p)$.
2) Find the poles of φ_h and φ_ℓ from the requirements in the attenuation-bands.
3) Modify the zeros of φ_h and φ_ℓ so that $|\varphi_h|^2$ and $|\varphi_\ell|^2$ do not exceed $1/G(p)$ and are tangent to $1/G(p)$ in their respective bands.
   Fig.-3 shows $|\varphi_h|^2$ and $|\varphi_\ell|^2$.
4) Obtain y_h, y_ℓ which satisfy the residue matching requirement(see(17)).
5) Proceed the cascade synthesis of y_h, y_ℓ.

The characteristics obtained through these procedures are shown in Figs. 4, 5, 6 and 7. Figs. 4 and 5 show the even part of the input admittance and the return loss at each port, respectively.

Fig.-6 shows the insertion loss between ports 1 and 3, ports 1 and 4.

Fig.-7 shows the network configuration and element values.

V. Conclusion

A practical design method of a four-port directional filter without use of compensating networks is proposed.

The example is designed so as to have the same electrical performances achieved by Desoer's method.

In addition, we realized a directional filter in which an inductor is used as phase inversion transformer and filter element (see Fig - 7). According to these results, the number of necessary circuit elements of the directional filter can be reduced.

VI. Acknowledgement

The author wishes to express his thanks to Mr. Tohru Tsuchiya and Mr. Yasutoshi Ishizaki, the managers of transmission division of Nippon Electric Co., for their kind suggestions and discussious.

References

[1] C. A. Desoer "Four-Port Constant R Network Required by a Two-Way Single Amplifier Repeater" IRE Trans on Circuit Theory December 1958 pp 276 ~ 283.

[2] Tohru Tsuchiya and Shun Nakao "On a Synthesis of Separating Filters" Proc of Kyoto Int. Conf on Circuit and System Theory, Sept. 9 ~ 11 1970.

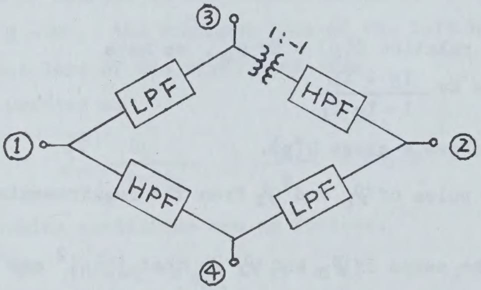


Fig.- 1 The configuration of the four-port directional filter

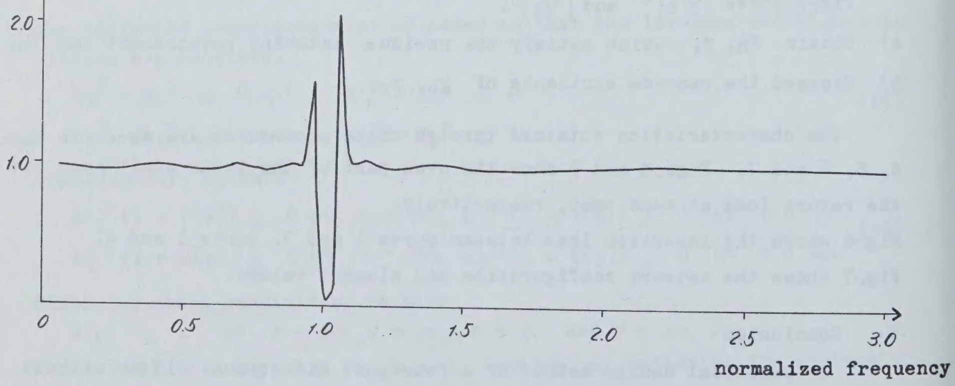


Fig.- 2 The characteristics of given G(p) (see eq(19))

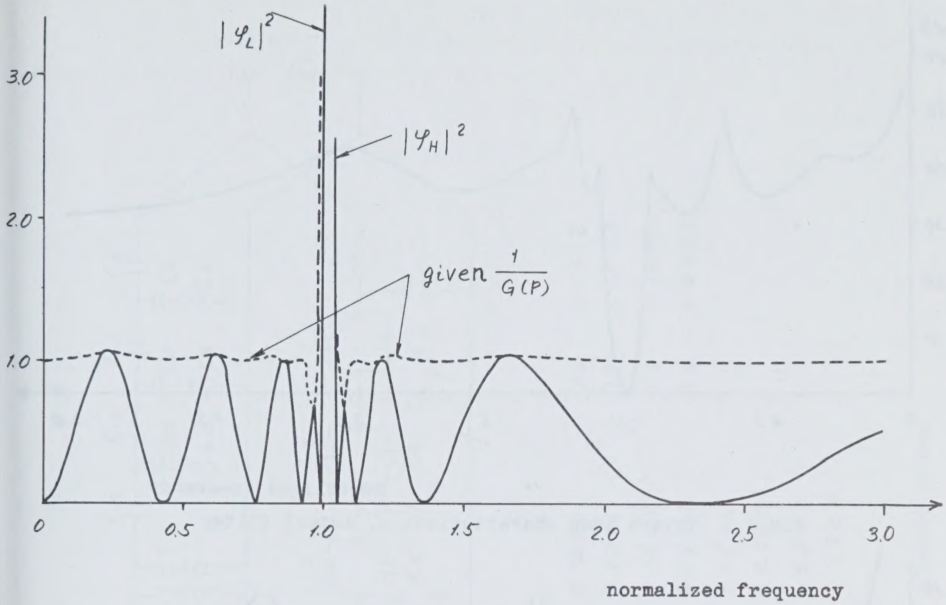


Fig.- 3 $|\varphi_H|^2$ and $|\varphi_L|^2$ after the design procedure 3).

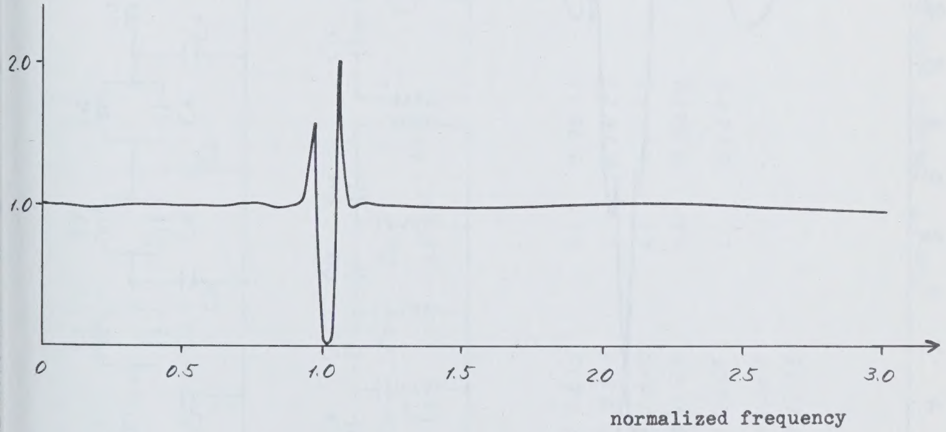


Fig.- 4 Even part of the input admittance characteristics, actual filter

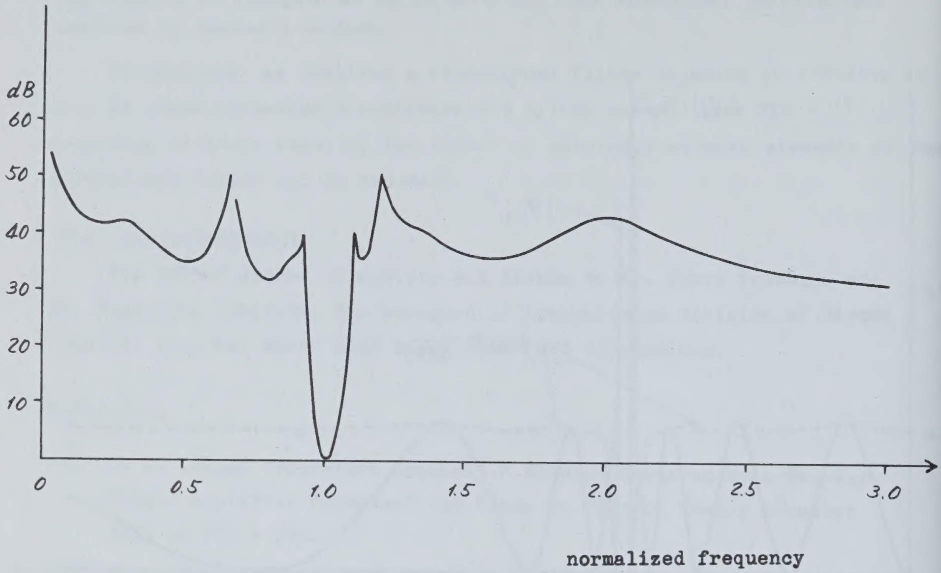


Fig.- 5 Return loss characteristics, actual filter

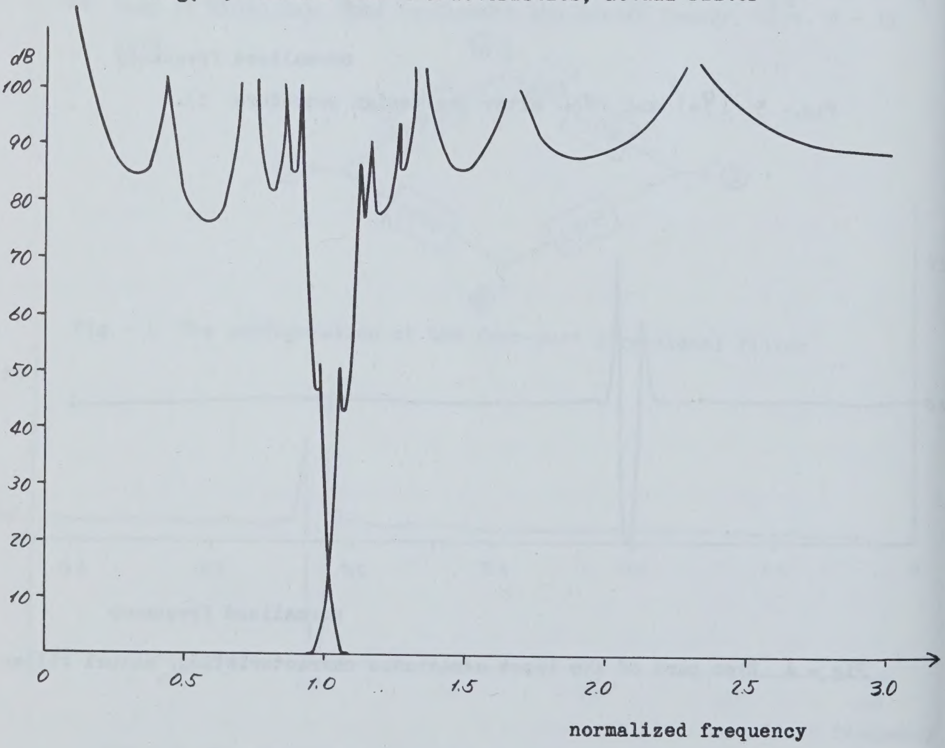
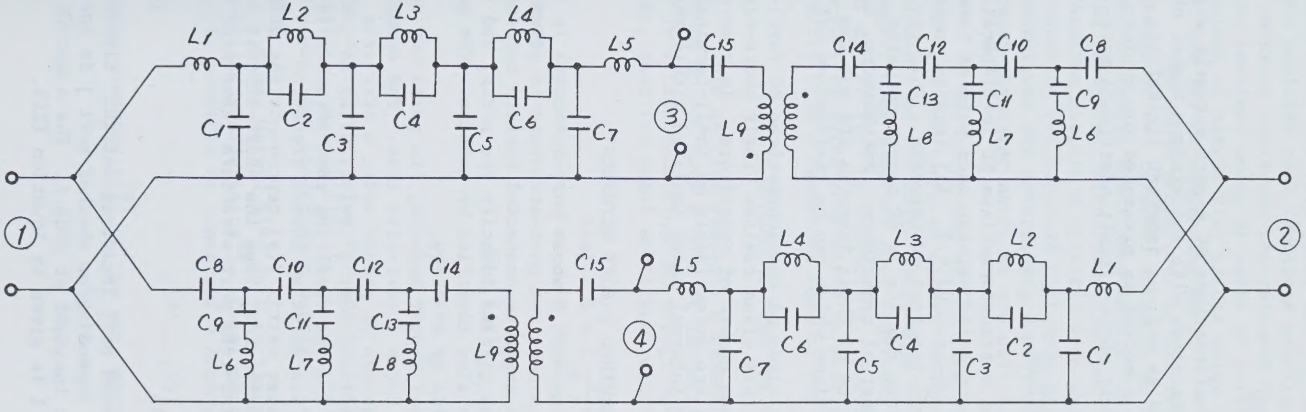


Fig.- 6 Insertion loss characteristics between ports 1 and 3, ports 1 and 4, actual filter


①
②
③
④
L1
L2
L3
L4
L5
L6
L7
L8
L9
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15

C1	0.2415	L1	0.2495	C8	0.1400	L6	0.1562
C2	0.0815	L2	0.1927	C9	0.2216	L7	0.1586
C3	0.1810	L3	0.1652	C10	0.1429	L8	0.1087
C4	0.1148	L4	0.2494	C11	0.1950	L9	0.0832
C5	0.2055	L5	0.2495	C12	0.1202		
C6	0.0367			C13	0.4704		
C7	0.2635			C14	0.0872		
				C15	0.0968		

Fig. - 7 Network configuration and element values

MINIMAL-CAPACITOR SYNTHESIS OF TRANSFORMERLESS RC n-PORT NETWORKS ON ONE-TREE PORT-STRUCTURES

DEAN J. SCHMIDLIN AND FRANK J. LUPO
New York University, Department of Electrical Engineering
Bronx, New York 10453

ABSTRACT

This paper presents a class of transformerless RC n-port networks which realize their short-circuit admittance matrices with the minimum number of capacitors. Some interesting properties as well as inherent limitations of these networks are discussed. Finally, a method is presented for synthesizing these RC n-port networks from their short-circuit admittance parameter matrices.

I. INTRODUCTION

Present techniques for synthesizing a transformerless RC (n+1)-terminal n-port network N generally assume that connected between each pair of terminals of network N is a two-terminal RC subnetwork[3,4,5]. All internal nodes of network N are thus isolated within two-terminal subnetworks. This restriction on the location of internal nodes results in RC n-port networks which contain more than the minimum number of capacitors. Furthermore, the short-circuit admittance matrices realized by these networks are not the most general class of RC admittance matrices which can be realized on n+1 port-terminals.

Presented in this paper is another class of transformerless RC (n+1)-terminal n-port networks. Networks in this class realize their short-circuit admittance matrices with the minimum number of capacitors. In general, these short-circuit admittance matrices are unrealizable as (n+1)-terminal n-port networks which have all of their internal nodes isolated within two-terminal subnetworks.

II. TRANSFORMERLESS (n+1)-TERMINAL n-PORT NETWORKS

Consider a transformerless n-port network N whose port-structure is in the form of a one-tree. An example of a one-tree port-structure is given in Fig. 1. The large circles and the line segments connected thereto represent ports. The numbers contained within the circles identify the ports, and the dots represent terminal nodes which are also identfied by numbers. The positive terminal of each port is identified by an arrow.

Since network N is an (n+1)-terminal n-port network, its ports may be redefined so that the new set of ports forms a star-like tree. For example, if node 0 in Fig. 1 represents the reference or ground node, a star-tree port-structure may be generated by redefining port i (i=1,2,...,n) as the port connected between node 0 and node i. The resulting port-strucure is shown in Fig. 2 where all of the ports are directed toward ground.

If short-circuit admittance parameter matrix Y(s) represents the original port-tree structure defined over network N, then the short-circuit admittance matrix corresponding to the star-tree port-structure (generated by the above method) is given by

$$\hat{Y}(s) = A\,Y(s)A^t \tag{1}$$

where transformation matrix A is a reduced port incidence matrix. Element a_{ij} in matrix A is either +1, -1, or 0 depending on whether port j is incident away from, incident toward, or not incident at node i. The A matrix corresponding to the port-tree in Fig. 1 is given by Equation (21).

Except for the difference in their port-structures, both parameter matrices $Y(s)$ and $\hat{Y}(s)$ represent network N. Since matrix $\hat{Y}(s)$ corresponds to a star port-tree, $\hat{Y}(s)$ is identical to the nodal admittance matrix of network N. If all of the internal nodes of network N are contained within two-terminal subnetworks, network N can be synthesized from matrix $\hat{Y}(s)$ by inspection. For this case, all of the off-diagonal elements of matrix $\hat{Y}(s)$ are positive real functions prefixed by a minus sign. These positive real functions represent the values of the admittances that are connected between the terminal nodes of pairs of ports. The sum of the elements in any row or column of matrix $\hat{Y}(s)$ is the value of the p.r. admittance that is connected across its corresponding port.

The above properties of the transformed matrix $\hat{Y}(s)$ simplify the synthesis of n+1 terminal networks on arbitrary port-tree structures. Given a short-circuit admittance matrix $Y(s)$, one simply determines its corresponding port-tree, transforms that port-tree structure to a star, and then synthesizes network N directly from nodal admittance matrix $\hat{Y}(s)$. Finally, by restoring the original port-tree structure one realizes matrix $Y(s)$. The techniques for determining port-tree structures from short-circuit admittance parameter matrices are well known and may be found in the literature [1,3,4,5,9].

If the locations of the internal nodes of network N are not restricted, i.e. are not isolated within two-terminal subnetworks, then the off-diagonal elements of the nodal admittance matrix $\hat{Y}(s)$ will not, in general, be the negative of positive real functions of s. This precludes the realization of matrix $\hat{Y}(s)$ by inspection. However, for the class of (n+1)-terminal n-port networks considered in this paper, network N may be synthesized directly from a canonical form of its nodal admittance matrix.

III. A CANONICAL FORM FOR RC NODAL ADMITTANCE MATRICES

This section presents a canonical form for RC nodal admittance matrices which do not have a pole at infinity. If the (n+1)-terminal n-port network N is a transformerless RC n-port network with no natural frequency at infinity, then its nodal admittance matrix $\hat{Y}(s)$ can be expressed as

$$\hat{Y}(s) = \hat{G} + \sum_{m=1}^{p} \frac{s}{s+\sigma_m} \hat{K}^{(m)} \tag{2}$$

where

$$\hat{G} = \hat{Y}(s)\big|_{s=0} \tag{3a}$$

$$\hat{K}^{(m)} = \frac{s+\sigma_m}{s} \hat{Y}(s)\big|_{s\to -\sigma_m} \tag{3b}$$

The above expression for matrix $\hat{Y}(s)$ is called the residue matrix expansion of $\hat{Y}(s)$. Matrix $\hat{K}^{(m)}$ is the matrix of residues at the negative real pole $s = -\sigma_m$, and conductance matrix $\hat{G}$ is the value of matrix $\hat{Y}(s)$ at zero value of s.

Since matrix $\hat{K}^{(m)}$ is a residue matrix of a passive RC network, it is positive semidefinite[11]. Consequently, if the rank of matrix $\hat{K}^{(m)}$ is r_m, it is possible to construct a decomposition

$$\hat{K}^{(m)} = T^{(m)} \, T^{(m)t} \tag{4}$$

where elements $t_{ij}^{(m)}$ of the $n \times r_m$ matrix $T^{(m)}$ have the properties

$$t_{ij}^{(m)} = 0 \qquad i < j \tag{5a}$$

$$t_{jj}^{(m)} > 0 \qquad j = 1,2,\ldots,r_m \tag{5b}$$

The values of elements $t_{ij}^{(m)}$ for $i>j$ may be negative, zero, or positive. The elements of matrix $T^{(m)}$ which are not specified by Equation (5a) are given by the recursion formulas[2]

$$t_{jj}^{(m)} = \left[\hat{k}_{jj}^{(m)} - \sum_{k=1}^{j-1} t_{jk}^{(m)\,2}\right]^{1/2} \tag{6a}$$

$$t_{ij}^{(m)} = t_{jj}^{(m)-1}\left[\hat{k}_{ij}^{(m)} - \sum_{k=1}^{j-1} t_{ik}^{(m)} t_{jk}^{(m)}\right] \qquad i = j + 1,\ldots n \tag{6b}$$

The above elements are determined in the order $t_{11}^{(m)}, t_{21}^{(m)}, \ldots, t_{n1}^{(m)}, t_{22}^{(m)}, t_{32}^{(m)}, \ldots, t_{n2}^{(m)}, \ldots, t_{r_m r_m}^{(m)}, \ldots, t_{nr_m}^{(m)}$.

If matrix $Q^{(m)}$ is an arbitrary orthogonal matrix of order r_m, then the matrix

$$X^{(m)} = T^{(m)}Q^{(m)} \tag{7}$$

is also a decomposition of residue matrix $\hat{K}^{(m)}$. In fact, Equation (7) defines all possible $n \times r_m$ matrices $X^{(m)}$ such that $X^{(m)}X^{(m)t} = \hat{K}^{(m)}$[12]. If column matrix $X_j^{(m)}$ represents the j-th column of matrix $X^{(m)}$, then matrix $\hat{K}^{(m)}$ may be expressed as

$$\hat{K}^{(m)} = \sum_{j=1}^{r_m} X_j^{(m)} X_j^{(m)t} \tag{8}$$

Substituting Equation (8) into Equation (2) gives

$$\hat{Y}(s) = \hat{G} + \sum_{m=1}^{p} \sum_{j=1}^{r_m} \frac{s}{s+\sigma_m} X_j^{(m)} X_j^{(m)^t} \tag{9}$$

Equation (9) represents the canonical form for nodal admittance matrix $\hat{Y}(s)$. The frequency sensitive portion of matrix $\hat{Y}(s)$ has been expressed as a sum of $\sum_{m=1}^{p} r_m$ residue matrices of rank one, each prefixed by a p.r. function of the type $s/(s+\sigma)$. Since the decomposition of a particular residue matrix is not unique, the above canonical form of matrix $\hat{Y}(s)$ is not unique.

If two or more (n+1)-terminal n-port networks having identical port-tree structures are connected in parallel, then it is well-known that the short-circuit admittance matrix of the composite network is equal to the sum of the short-circuit admittance matrices of the component networks. Consequently, a possible approach to the realization of nodal admittance matrix $\hat{Y}(s)$ is to realize each rank one residue matrix as a separate network, and then connect all of the separate networks in parallel. The nodal admittance matrix of each component network and hence of the composite network will in general have a conductance matrix associated with it. Otherwise, the requirement that a nodal admittance matrix must be hyperdominant for all real positive values of s cannot be satisfied. Since the conductance matrix of the composite network will generally not be equal to $\hat{G}$, a resistive (n+1)-terminal network must be added in parallel.

A fundamental (n+1)-terminal network which realizes any RC nodal residue matrix of rank 1 is illustrated in Fig. 3. The network consists of a single capacitor which is connected between apex nodes a_1 and a_2, and n+1 resistors. k of the resistors are connected to apex node a_1 and n-k+1 resistors to apex node a_2. The nodal admittance matrix of the network with respect to node 0 is given by[10].

$$Y^*(s) = P + \frac{s}{s+\sigma} XX^t \tag{10}$$

where matrix P is equal to the direct sum[8] of the following two conductance matrices:

$$P_{11} = \mathrm{DIAG}[g_1, g_2, \ldots, g_k] - \frac{1}{g^{(1)}} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_k \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_k \end{bmatrix}^t \tag{11a}$$

$$P_{22} = \mathrm{DIAG}[g_{k+1}, g_{k+2}, \ldots, g_n] - \frac{1}{g^{(2)}} \begin{bmatrix} g_{k+1} \\ g_{k+2} \\ \vdots \\ g_n \end{bmatrix} \begin{bmatrix} g_{k+1} \\ g_{k+2} \\ \vdots \\ g_n \end{bmatrix}^t \tag{11b}$$

where

$$g^{(1)} = \sum_{i=1}^{k} g_i \tag{12a}$$

$$g^{(2)} = g_o + \sum_{i=k+1}^{n} g_i \tag{12b}$$

The entries x_j of column matrix X are given by

$$x_j = \alpha g_j \qquad j = 1,2,\ldots,k \tag{13a}$$

$$x_j = -\beta g_j \qquad j = k+1, k+2, \ldots, n \tag{13b}$$

where

$$\alpha^{-1} = \sqrt{g^{(1)}}\,(1+a^{-1})^{1/2} \tag{14a}$$

$$\beta^{-1} = \sqrt{g^{(2)}}\,(1+a)^{1/2} \tag{14b}$$

$$a = g^{(2)}/g^{(1)} \tag{14c}$$

The first k entries of X are nonnegative, whereas the last n-k entries are nonpositive. The value of σ is

$$\sigma = \frac{g^{(2)}/C}{1+a} \tag{15}$$

The design of the network in Fig. 3 is carried out by expressing the network elements in terms of σ,a, and the n entries of column matrix X. Let

$$x^{(1)} = \sum_{i=1}^{k} x_i \tag{16a}$$

$$x \;\;= \sum_{i=1}^{n} x_i \tag{16b}$$

The values of the network elements are given by

$$g_i = (1+a^{-1})\, x^{(1)} x_i \qquad i = 1,2,\ldots,k \tag{17a}$$

$$g_i = -(1+a)\, x^{(1)} x_i \qquad i = k+1,k+2,\ldots,n \tag{17b}$$

$$g_o = (1+a)\, x^{(1)}\, x \tag{17c}$$

$$C = x^{(1)^2}\, \sigma^{-1} \tag{17d}$$

From Equation (17c) one sees that the column sum x of column matrix X must be nonnegative. However, if the given x is negative, then reversing the sign of each element of matrix X generates a positive x. The modifications in sign do not affect the product XX^t. Because of the node numbering assigned to the terminal nodes of the network prototype, the first k elements of X are nonnegative, and the last n-k elements are nonpositive. A different distribution of the positive and negative elements of X correspond to a different numbering of the terminal nodes (nonzero conductances which are connected to apex node a_1 correspond to positive elements in X, whereas nonzero conductances connected to apex node a_2 correspond to negative elements in X). The value of the nonnegative parameter a defined by Equation (14c) is arbitrary. Once the values of the conductances g_i $(i=0,1,2,\ldots,n)$ are computed, the values of conductance matrices P_{11} and P_{22} may be determined from Equations (11).

The network prototype of Fig. 3 is capable of realizing **any** RC nodal residue matrix (at a finite pole) of rank unity. Let matrix $P_j^{(m)}$ be the nodal conductance matrix associated with the realization of matrix $X_j^{(m)}X_j^{(m)t}$. The nodal admittance matrix of the composite network N_f which realizes the frequency sensitive component of matrix $\hat{Y}(s)$ is

$$\hat{Y}_f(s) = P_f + \sum_{m=1}^{p} \sum_{j=1}^{r_m} \frac{s}{s+\sigma_m} X_j^{(m)} X_j^{(m)t} \tag{18a}$$

where

$$P_f = \sum_{m=1}^{p} \sum_{j=1}^{r_m} P_j^{(m)} \tag{18b}$$

It is always possible to realize the frequency sensitive component of matrix $\hat{Y}(s)$ by a parallel combination of networks similar to the network prototype of Fig. 3. However, conductance matrix P_f may not be compatible with conductance matrix $\hat{G}$. For this reason, matrices P_f and $P_j^{(m)}$ are called parasitic conductance matrices.[6,7] A sufficient condition for the realization of matrix $\hat{Y}(s)$ is that matrix

$$\hat{G}_r = \hat{G} - P_f \tag{19}$$

be hyperdominant. If $\hat{G}_r$ is hyperdominant, then it corresponds to a resistive (n+1)-terminal network N_r which when added in parallel with network N_f realizes nodal admittance matrix $\hat{Y}(s)$.

IV. A SYNTHESIS PROCEDURE

This section presents a procedure for realizing RC short-circuit admittance matrix Y(s) on n+1 terminals. The procedure consists of the five following parts:

(1) Determination of the port-tree structure which corresponds to matrix Y(s).

(2) Transformation of the port-structure to a star port-tree and determination of the nodal admittance matrix $\hat{Y}(s)$.
(3) Decomposition of matrix $\hat{Y}(s)$ into canonical form.
(4) Synthesis of network $N=N_r||N_f$ from the canonical form of matrix $\hat{Y}(s)$.
(5) Recovery of the port-tree structure determined in Part (1).

The first part of the procedure is accomplished by applying any of the well-known port-tree synthesis methods[1,3,4,5,9]. The second part of the procedure, the transformation of port-structure and determination of matrix $\hat{Y}(s)$ is discussed in Section II of this paper. In Section III, methods are discussed for expressing matrix $\hat{Y}(s)$ in canonical form and synthesing network N from this form. If the port-tree structure determined in Part (1) is recovered, the resulting (n+1)-terminal n-port network realizes matrix Y(s).

The most crucial step of the synthesis procedure is the generation of a parasitic conductance matrix P_f which is compatible with matrix $\hat{G}$ (that is, a matrix P_f such that $\hat{G}-P_f$ is hyperdominant). Each set of orthogonal matrices $Q^{(m)}(m=1,2,\ldots,p)$ defines a set of decomposition matrices $X^{(m)}(m=1,2,\ldots,p)$ which in turn defines a parasitic conductance matrix P_f. Currently, work is being done to minimize the effects of P_f by an appropriate choice of the orthogonal matrices $Q^{(1)},Q^{(2)},\ldots,Q^{(p)}$. The fundamental network of Fig. 3 can be made more practical by placing a conductance in parallel with capacitor C. This conductance can account for losses occurring in the dielectric of non-ideal capacitors. However, this conductance also increases the parasitic effects of conductance matrix P_f[10].

The total number of Rank 1 residue matrices in the canonical form of matrix $\hat{Y}(s)$ (which is also the total number of capacitors in network N) is equal to the sum of the ranks of the residue matrices of matrix Y(s). Since this number is the minimum number of capacitors needed for a network realization of an RC admittance matrix[8], network N represents a minimal-capacitor realization of Y(s).

In Section III it was assumed that nodal admittance matrix $\hat{Y}(s)$ is finite at s equal to infinity. If matrix $\hat{Y}(s)$ does posses a pole at infinity, then the nodal capacitance matrix $\hat{C}$ will be nonzero and hyperdominant. Consequently, $\hat{C}$ can be realized by the method discussed in Section II. However, the network realizing $\hat{C}$ will contain as many as $n(n+1)/2$ capacitors, which is more than the minimum number, rank $\hat{C}$. The above network is an example of an (n+1)-terminal network which has all of its internal nodes isolated within two-terminal subnetworks (in the case of $\hat{C}$, no internal nodes). All residue matrices of the nodal admittance matrix of such a network must be hyperdominant. For the class of n-port networks presented in this paper, the nodal residue matrices need only be positive semidefinite.

Example: Consider the RC short-circuit admittance matrix

$$Y(s) = \begin{bmatrix} 49 & 45 & 41 & 8 \\ 45 & 121 & 107 & 62 \\ 41 & 107 & 133 & 68 \\ 8 & 62 & 68 & 83 \end{bmatrix} + \frac{s}{s+10}\begin{bmatrix} 5 & -5 & -9 & 7 \\ -5 & 10 & 1 & -18 \\ -9 & 1 & 29 & 5 \\ 7 & -18 & 5 & 34 \end{bmatrix} \tag{20}$$

The short-circuit conductance matrix $G=Y(0)$ is uniformly tapered[3,4,5]. Consequently, the port-tree structure of network N is the linear port-tree illustrated in Fig. 1. The reduced incidence matrix A of this port-tree is

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{21}$$

The value of nodal admittance matrix $\hat{Y}(s)$ is

$$\hat{Y}(s) = \begin{bmatrix} 80 & -10 & -12 & -54 \\ -10 & 40 & -20 & -6 \\ -12 & -20 & 80 & -15 \\ -54 & -6 & -15 & 83 \end{bmatrix} + \frac{s}{s+10} \begin{bmatrix} 25 & -5 & -35 & 25 \\ -5 & 37 & -5 & -23 \\ -35 & -5 & 53 & -29 \\ 25 & -23 & -29 & 34 \end{bmatrix} \tag{22}$$

Matrix $\hat{Y}(s)$ has a single residue matrix $\hat{K}^{(1)}$. Computing matrix $T^{(1)}$ from Equations (6) gives

$$T^{(1)} = \begin{bmatrix} 5 & 0 \\ -1 & 6 \\ -7 & -2 \\ 5 & -3 \end{bmatrix} \tag{23}$$

Letting orthogonal matrix $Q^{(1)}$ be equal to the identity matrix results in $X^{(1)}=T^{(1)}$. A network which realizes the residue matrix $X_1^{(1)}X_1^{(1)t}$ has the element values

$$\begin{array}{lll} g_o = 40 & g_2 = 20 & g_4 = 100 \\ g_1 = 100 & g_3 = 140 & C = 10 \end{array} \tag{24}$$

The above element values were computed from Equations (17) with parameter a equal to unity. If the i-th entry of $X_1^{(1)}$ is positive, then g_i is computed from Equation (17a). If the i-th entry of $X_1^{(1)}$ is negative, then g_i is computed from Equation (17b). The value of $x^{(1)}$ is the sum of the positive elements of $X_1^{(1)}$. A network which realizes residue matrix $X_2^{(1)}X_2^{(1)t}$ has the element values

$$\begin{array}{lll} g_o = 12 & g_2 = 72 & g_4 = 36 \\ g_1 = 0 & g_3 = 24 & C = 3.6 \end{array} \tag{25}$$

The parasitic conductance matrices associated with the above networks are

$$P_1^{(1)} = \begin{bmatrix} 50 & 0 & 0 & -50 \\ 0 & 18 & -14 & 0 \\ 0 & -14 & 42 & 0 \\ -50 & 0 & 0 & 50 \end{bmatrix} \tag{26a}$$

$$P_2^{(1)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & -12 \\ 0 & 0 & -12 & 18 \end{bmatrix} \tag{26b}$$

The composite parasitic conductance matrix is

$$P_f = P_1^{(1)} + P_2^{(1)} = \begin{bmatrix} 50 & 0 & 0 & -50 \\ 0 & 18 & -14 & 0 \\ 0 & -14 & 58 & -12 \\ -50 & 0 & -12 & 68 \end{bmatrix} \tag{27}$$

Computing matrix $\hat{G} - P_f$ gives

$$\hat{G}_r = \begin{bmatrix} 30 & -10 & -12 & -4 \\ -10 & 22 & -6 & -6 \\ -12 & -6 & 22 & -3 \\ -4 & -6 & -3 & 15 \end{bmatrix} \tag{28}$$

Equations (24), (25), and (28) define a network N which realizes Matrix (20). This network is illustrated in Fig. 4. Equivalent networks may be generated by varying either the orthogonal matrix $Q^{(1)}$ or the a parameters of the two component subnetworks of network N_f.

V. CONCLUSION

This paper has presented a class of transformerless RC (n+1)-terminal n-port networks which realize their short-circuit admittance matrices Y(s) with the minimum number of capacitors. The residue matrices of admittance matrix Y(s) have only the general property of being positive semidefinite (all residue matrices of passive networks must be positive semidefinite). Some inherent limitations of the above class of n-port networks are: (1) the short-circuit admittance matrix Y(s) of this class of networks cannot have a pole at infinity, and (2) admittance matrix Y(s) cannot be zero at s=0.

Presently, attention is being given to the problem of formulating systematized procedures for (1) the optimization of parasitic conductance matrix P_f, and (2) the synthesis of transformerless RL and RLC (n+1)-terminal n-port networks with the minimum number of reactive elements.[10]

ACKNOWLEDGMENT

The authors are indebted to the National Science Foundation for their partial support of this work (NSF Grant GK-2948).

REFERENCES

1. F. Boesch, "On the Synthesis of Resistor N-Ports," Ph.D. Diss., Polytechnic Institute of Brooklyn, N.Y., 1963.
2. J.N. Franklin, Matrix Theory, Prentice-Hall, 1968.
3. E.A. Guillemin, Theory of Linear Physical Systems, John Wiley and Sons, Inc., New York, New York, 1963.
4. C.C. Halkias, "Analysis and Synthesis of N-Port Networks," Ph.D. Diss., Columbia University, N.Y., 1962.
5. W.H. Kim and R.T. Chien, Topological Analysis and Synthesis of Communication Networks, Columbia University Press, New York, 1962.
6. F.J. Lupo and C.C. Halkias, "Synthesis of n-Port Networks on Two-Tree Port-Structures," IEEE Trans. Circuit Theory, Vol. CT-12, pp. 571-577, December 1965.
7. F.J. Lupo, "The Synthesis of Transformerless n-Port Networks on Multi-tree Port Structures," IEEE Trans. Circuit Theory, Vol. CT-15, pp. 211-220, September 1968.
8. R.W. Newcomb, Linear Multiport Synthesis, McGraw-Hill, New York, 1966.
9. D.J. Schmidlin and F.J. Lupo, "The Determination of Port-Tree Structures of Transformerless (n+1)-Terminal n-Port Networks from their Short-Circuit Admittance Parameter Matrices," Proc. Fifth Annual Allerton Conference on Circuit and System Theory, pp. 748-756, University of Illinois, October 1967.
10. D.J. Schmidlin and F.J. Lupo, "The Synthesis of Transformerless (n+1)-Terminal n-Port Networks with the Minimum Number of Reactive Elements," New York University Technical Report (in preparation).
11. L. Weinberg and P. Slepian, "Realizability Conditions on n-Port Networks," IRE Transactions on Circuit Theory, Vol. CT-5, pp. 217-221, September 1958.
12. B. Vinograde, "Canonical Positive Definite Matrices Under Internal Linear Transformations," Proc. Am. Math. Soc., Vol. 1, pp. 159-161, 1950.

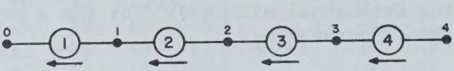


FIG. 1

A LINEAR PORT-TREE

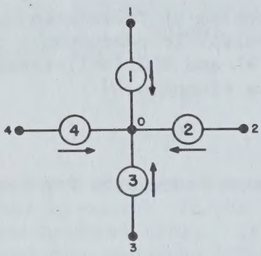


FIG. 2

A STAR PORT-TREE

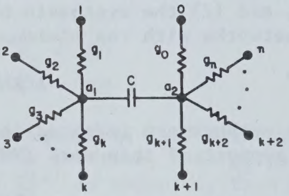

1
0
2
g_2
g_1
g_0
g_n
n
a_1
C
a_2
g_3
3
g_k
k
g_{k+1}
k+1
g_{k+2}
k+2

FIG. 3

A FUNDAMENTAL RC (n+1) - TERMINAL NETWORK

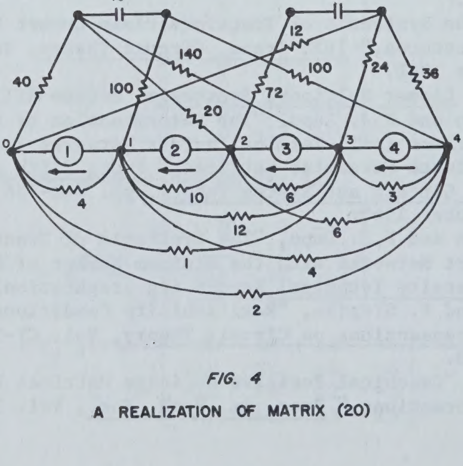

10
3.6
12
140
100
24
36
40
100
72
20
0
1
2
3
4
1
2
3
4
4
10
6
3
12
6
1
4
2

FIG. 4

A REALIZATION OF MATRIX (20)

AN ALGORITHM FOR THE IDENTIFICATION AND AUGMENTATION OF RESISTIVE IRREGULAR NETWORKS

J. B. OSO, T. E. STERN, and H. E. MEADOWS
Systems Research Group, Columbia University
New York, N.Y. 10027

Abstract — This paper develops an easily programmed algorithm which identifies a resistive irregular network and augments the network accordingly so that the resulting network can be characterized in a canonical form. The elements admitted into the network consist of quasilinear capacitors, quasilinear inductors, quasilinear resistors, voltage-controlled negative resistors and current-controlled negative resistors.

I. INTRODUCTION

Many authors [1,2,3,4] have done work in regard to the characterization of special nonlinear networks in the canonical form represented by

$$\dot{\underline{x}} = \underline{f}(\underline{x},t) \quad . \quad . \quad . \qquad (1)$$

where $\underline{x}$ is a vector representing the state variables.

Stern, Oh, and Meadows [5] have proposed a systematic procedure for characterizing a broad class of nonlinear networks in the canonical form. In [5], the concept of the regularity of a network was introduced. A set of state variables $(\underline{q},\underline{\lambda})$ is chosen based on a proper tree T_N of the network selected as outlined in [5]. If the selected set $(\underline{q},\underline{\lambda})$ constitutes a complete set of state variables, the network is defined as regular and can be represented in the canonical form. Otherwise, the network is defined to be strictly irregular. If the network topology and the network elements characterization are not sufficient to determine regularity of the network, the network is classified as potentially irregular. If a network is irregular only because of the topology or the characterization of the resistive elements of the network, the network is called resistive irregular. The definitions of a quasilinear element, voltage-controlled negative resistor (VCNR), and current-controlled negative resistor (CCNR) are also in [5].

This paper is based on [5] and is concerned with developing an easily programmed algorithm which identifies a resistive irregular network and augments the network accordingly so that the resulting network can be characterized in the canonical form. The elements admitted into the network consist of quasilinear capacitors (QLC), quasilinear resistors (QLR), current-controlled negative resistors (CCNR), voltage-controlled negative resistors (VCNR), and quasilinear inductors (QLL).

II. IDENTIFICATION OF RESISTIVE IRREGULAR NETWORKS

In [6] Oh proved the following theorem which identifies a resistive irregular network.

Theorem

Given a network N in the class defined above, form a resistive subnetwork N_R by short-circuiting capacitive elements and open-circuiting inductive elements. The network N is strictly irregular if the resulting resistive subnetwork N_R contains any VCNR open-branches or CCNR self-loops. The network is potentially irregular if the subnetwork contains a nonseparable connected resistive subnetwork of two or more elements containing VCNR or CCNR.

III. AUGMENTATION OF RESISTIVE IRREGULAR NETWORKS

If the network N is irregular as stated in the theorem, the following augmentation is proposed.

1. Form a complete forest of the resistive subnetwork N_R using elements in the order of VCNR, QLR, CCNR.

2. With respect to this particular choice of a forest, add parasitic capacitors ($C = \in$) in parallel with every VCNR forest branch, and add parasitic inductors ($L = \in$) in series with every CCNR chord or CCNR self-loops of N_R.

If the modified network is denoted as N_M, it is shown [6] that N_M can be characterized in a canonical form by choosing state variables as in [5]; that is, that N_M is regular.

IV. ALGORITHM FOR IDENTIFYING RESISTIVE IRREGULAR NETWORKS

For large networks, it is very convenient that the identification and augmentation of an irregular network be done by computer. This section gives an easily programmed algorithm for the purpose of identification.

Suppose a network N is given with

n = total number of nodes of N

ℓ = total number of QLL elements of N

c = total number of QLC elements of N

b = total number of branches of N

b_R = $b - \ell - c$ = total number of resistive elements of N

A_{nxb} = complete node-cutset matrix of N with the columns of A arranged in the order: VCNR, QLR, CCNR, QLC, QLL

S_ℓ = Set of QLL elements of N

S_c = Set of QLC elements of N

S_r = Set of resistive elements of N

Consider for example, the network N_e in Fig. 1 with the nodes numbered from a to f and the branches numbered from 1 to 9.

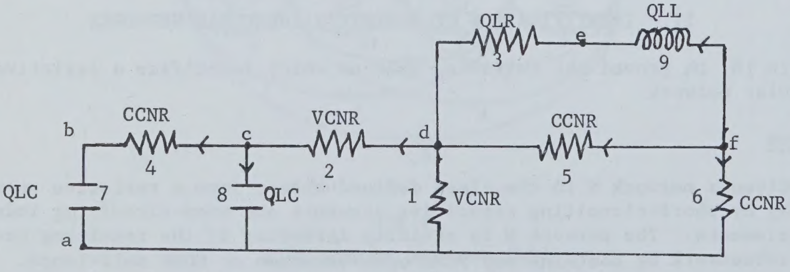


Fig. 1. A given network N_e.n=6,c=2,b_R=6, ℓ=1

The complete node-cutset matrix A_e for this network is given by

$$A_e = \left[\begin{array}{c|r r@{\ :\ }r@{\ :\ }r r r@{\ :\ }r r@{\ :\ }r}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline
a & -1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 \\
b & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\
c & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
d & 1 & 1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
e & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\
f & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1
\end{array}\right]$$

$$\text{VCNR} \quad \text{QLR} \quad \text{CCNR}$$

$$\longleftarrow S_r \longrightarrow \longleftarrow S_c \longrightarrow S_\ell$$

The proposed algorithm consists of operations on A_e as outlined below.

Step 1 Forming resistive subnetwork N_R

(a) Open-circuiting the QLL elements does not change the number of rows of the node-cutset matrix. Therefore form the matrix $Q'_{n\times(b-\ell)}$ from A_e by neglecting the last ℓ columns corresponding to S_ℓ.

$$Q' = \left[\begin{array}{c|r r@{\ :\ }r@{\ :\ }r r r@{\ :\ }r r}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\hline
a & -1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 \\
b & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\
c & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 1 \\
d & 1 & 1 & -1 & 0 & -1 & 0 & 0 & 0 \\
e & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
f & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0
\end{array}\right]$$

(b) For each capacitor shorted a node is destroyed. Therefore shorting the c elements of S_c reduces Q' to $Q''(n-c)x(b- -c)$. Q'' is derived from Q' as follows:

For each column S_{ci} ($i = 1,2, \cdots c$) of S_c, locate the position of the +1 and the -1, add the rows, say i and j, in which these are located, replace the two rows i and j by the single row i+j representing the sum. Cancel the corresponding column S_{ci} of Q'. Each time this process is carried out, the number of columns is reduced by 1 and the number of rows is reduced by 1. Continue the process until all the columns corresponding to S_c have been eliminated. The remaining $(n-c)x(b-\ell-c)$ matrix is Q''. Cancel the last row of Q'' to get Q, the incidence matrix of the resistive subnetwork N_R of the network N. If any column of zeros turns up, the corresponding edge is a self-loop. Hence, edge 4 is a self-loop of N_R. This is noted and the edge is removed before Step 2 begins.

$$Q''_8 \quad \begin{array}{c|ccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline a+c & -1 & -1 : & 0 : & 1 & 0 & -1 : & -1 \\ b & 0 & 0 : & 0 : & -1 & 0 & 0 : & 1 \\ d & 1 & 1 : & -1 : & 0 & -1 & 0 : & 0 \\ e & 0 & 0 : & 1 : & 0 & 0 & 0 : & 0 \\ f & 0 & 0 : & 0 : & 0 & 1 & 1 : & 0 \end{array}$$

QLC 8 is short-circuited

$$Q''_7 = \begin{array}{c|cccccc|} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline (a+c)+b & -1 & -1 : & 0 : & 0 & 0 & -1 \\ d & 1 & 1 : & -1 : & 0 & -1 & 0 \\ e & 0 & 0 : & 1 : & 0 & 0 & 0 \\ f & 0 & 0 : & 0 : & 0 & 1 & 1 \end{array}$$

QLC 7 is short-circuited

$$Q = \begin{array}{c|ccccc|} & 1 & 2 & 3 & 5 & 6 \\ \hline (a+c)+b & -1 & -1 : & 0 : & 0 & -1 \\ d & 1 & 1 : & -1 : & -1 & 0 \\ e & 0 & 0 : & 1 : & 0 & 0 \end{array}$$

f is chosen as reference.

The self-loop 4 of N_R is removed.

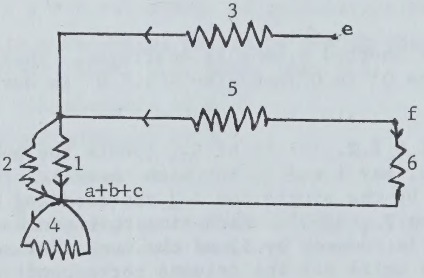


Fig. 2. Resistive subnet N_R of network N

Step 2 Forming a complete forest of N_R

With respect to N_R the KCL equation is

$$\underline{Q}\ \underline{i} = 0 \quad \cdot \quad \cdot \quad \cdot \qquad (2)$$

Let n_R = number of rows of Q.

The columns of Q are still arranged in the order of VCNR, QLR, CCNR. The next step consists of performing row operations on Q in an attempt to reduce (2) to the form [7]:

$$[-I \vdots F] \underline{i} = \underline{0} \quad . \quad . \quad . \qquad (3)$$

F is the familiar fundamental loop matrix of the circuit N_R with respect to the tree defined by the first (n_R-1) columns. When this reduction is attempted it may not be possible to obtain the desired result because the branches represented by the first n_R-1 columns of Q do not form a tree of N_R. In any case the matrix should be reduced to row echelon form. Each branch corresponding to the column with the first -1 in each row is a tree branch of N_R. The totality of such branches form the forest of N_R. The row-reduced matrix may then be brought into the desired form by a transposition of columns. If zero rows turn up at the bottom of F during reduction, the graph of the network N_R is not connected. Reduction of the matrix Q above results in the coefficient matrix Q_c.

$$Q_c = \begin{array}{c|ccccc} & 1 & 2 & 3 & 5 & 6 \\ \hline g & -1 & -1 & 0 & 0 & -1 \\ d & 0 & 0 & -1 & -1 & -1 \\ e & 0 & 0 & 0 & -1 & -1 \end{array}$$

The proper tree of N_R thus consists of edge 1,3,5, as shown in Fig. 2.

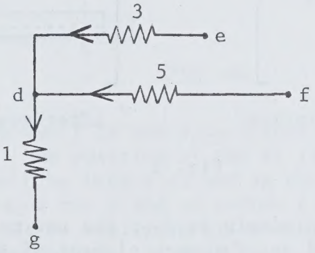


Fig. 2. Proper tree of N_R

Step 3 Identification

Let

S_{vc} = the set of VCNR forest branches of N_R

S_{cc} = the set of CCNR chords (including self-loops).

$$S_{vc} = \emptyset, \text{ the empty set} \quad . \quad . \quad . \qquad (4)$$

and

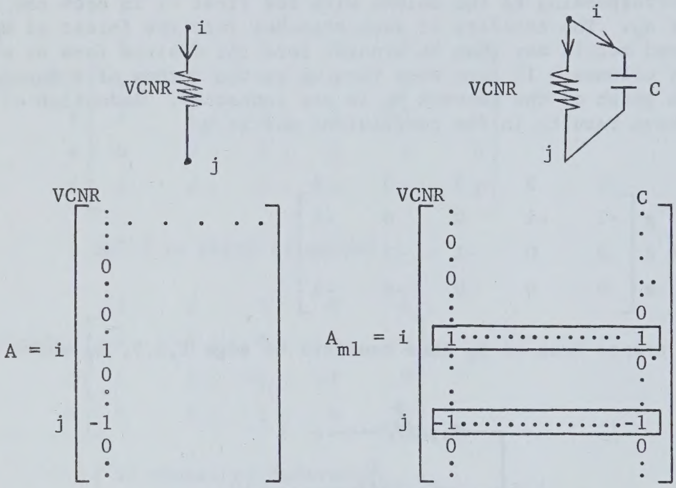
$$S_{cc} = \emptyset \quad . \quad . \quad . \quad . \quad . \qquad (5)$$

then the original network N is regular. If either of (4) or (5) is not satisfied, the network N is irregular. In the latter case augmentation is accomplished as follows.

V. ALGORITHM FOR AUGMENTING RESISTIVE IRREGULAR NETWORKS

We now go back to the original matrix A.

(a) For each element in S_{vc} a new capacitive branch ($C = \epsilon$) is defined. This adds a new column c to the set S_c of A without changing the number of rows since no new node is added. The location of the +1 and the -1 in the column c is identical with the same rows as in the column corresponding to the branch for which it was defined. See Fig. 3.



Before augmentation After augmentation

Fig. 3

If p is the total number of elements in S_{vc}, the new matrix A_{m1}, resulting after this process is carried out for each element of S_{vc}, is of the order $n \times (b+p)$. For the network N_e in Fig. 1, $S_{vc} = \{1\}$. Therefore the matrix A_{m1} is given by

$$A_{m1} = \begin{array}{c|cccccccccccccc}
 & 1 & 2 & & 3 & & 4 & 5 & 6 & & 7 & 8 & c & & 9 \\
\hline
a & -1 & 0 & : & 0 & : & 0 & 0 & -1 & : & -1 & -1 & -1 & : & 0 \\
b & 0 & 0 & : & 0 & : & -1 & 0 & 0 & : & 1 & 0 & 0 & : & 0 \\
c & 0 & -1 & : & 0 & : & 1 & 0 & 0 & : & 0 & 1 & 0 & : & 0 \\
d & 1 & 1 & : & -1 & : & 0 & -1 & 0 & : & 0 & 0 & 1 & : & 0 \\
e & 0 & 0 & : & 1 & : & 0 & 0 & 0 & : & 0 & 0 & 0 & : & -1 \\
f & 0 & 0 & : & 0 & : & 0 & 1 & 1 & : & 0 & 0 & 0 & : & 1
\end{array}$$

(b) For each element S_{cci} in S_{cc} define a new inductive branch ($L = \epsilon$) and a new node. That is, for each element of S_{cc} a new row and a new column are added to the matrix A_{m1}. This addition if effected as follows.

Suppose that the element S_{cci} is connected between the nodes i and j, represented by the presence of a +1 and a -1 in the i and j rows respectively of the S_{cci} column of A_{m1}. Denote the new column by L and the new row by g. See Fig. 4.

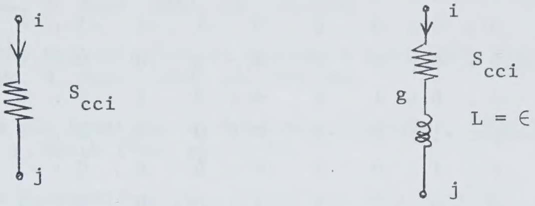


Fig. 4a

$$A_{m1} = \begin{array}{c} \\ \\ i \\ \\ \\ j \\ \\ \end{array} \overset{S_{cci}}{\left[\begin{array}{cccc} 0 & & & \\ \vdots & & & \\ 1 & \cdot & \cdot\;\cdot & \cdot \\ 0 & & & \\ \vdots & & & \\ -1 & & & \\ 0 & & & \\ \vdots & & & \end{array}\right]}$$

Before augmentation

$$A_{m2} = \begin{array}{c} \\ \\ i \\ \\ \\ j \\ \\ g \\ \\ \end{array} \left[\begin{array}{ccc} S_{cci} & & L \\ 0 & & \vdots \\ \vdots & & 0 \\ \hline 1 & \cdots\cdots & 0 \\ \hline 0 & & \vdots \\ \vdots & & 0 \\ \hline 0 & \cdots\cdots & -1 \\ \hline 0 & & 0 \\ \vdots & & \vdots \\ \hline -1 & \cdots\cdots & +1 \\ \hline 0 & & 0 \\ \vdots & & \vdots \end{array}\right]$$

After augmentation

Fig. 4b

Replace the -1 in row j in the S_{cci} column by 0 and write -1 in the row g of the column S_{cci}. The position of the +1 is unchanged in column S_{cci}. In the row j of column L we have a -1 and in the row g we have a +1. All the remaining elements of row g and of column L are 0. The matrix A_{m1} is otherwise unchanged. It is easily verified that these matrix operations effect the required augmentation. Suppose there are q elements in the set S_{cc}. With the process carried out for each element of S_{cc}, the resulting matrix D, of the order (n+q)x(b+p+q), is the complete node cutset matrix of the augmented network. By the theorem stated before, D together with the branch relations of the b+p+q elements is enough to put the modified network in the canonical form.

$$S_{cc} = \{4,6\}$$

$$A_{m2} = \begin{array}{c|ccccccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & c & 9 & L_1 \\ \hline a & -1 & 0\,: & 0\,: & 0 & 0 & -1\,: & -1 & -1 & -1\,: & 0 & 0 \\ b & 0 & 0\,: & 0\,: & 0 & 0 & 0\,: & 1 & 0 & 0\,: & 0 & -0 \\ c & 0 & -1\,: & 0\,: & 1 & 0 & 0\,: & 0 & 1 & 0\,: & 0 & 0 \\ d & 1 & 1\,: & -1\,: & 0 & -1 & 0\,: & 0 & 0 & 1\,: & 0 & 0 \\ e & 0 & 0\,: & 1\,: & 0 & 0 & 0\,: & 0 & 0 & 0\,: & -1 & 0 \\ f & 0 & 0\,: & 0\,: & 0 & 1 & 1\,: & 0 & 0 & 0\,: & 1 & 0 \\ g & 0 & 0\,: & 0\,: & -1 & 0 & 0\,: & 0 & 0 & 0\,: & 0 & 1 \end{array}$$

An inductive element L_1 is defined for edge 4

$$D = \begin{array}{c|cccccccccccccccc|}
 & 1 & 2 & & 3 & & 4 & 5 & 6 & & 7 & 8 & c & & 9 & L_1 & L_2 \\
a & -1 & 0 & : & 0 & : & 0 & 0 & 0 & : & -1 & -1 & -1 & : & 0 & 0 & -1 \\
b & 0 & 0 & : & 0 & : & 0 & 0 & 0 & : & 1 & 0 & 0 & : & 0 & -1 & 0 \\
c & 0 & -1 & : & 0 & : & 1 & 0 & 0 & : & 0 & 1 & 0 & : & 0 & 0 & 0 \\
d & 1 & 1 & : & -1 & : & 0 & -1 & 0 & : & 0 & 0 & 1 & : & 0 & 0 & 0 \\
e & 0 & 0 & : & 1 & : & 0 & 0 & 0 & : & 0 & 0 & 0 & : & -1 & 0 & 0 \\
f & 0 & 0 & : & 0 & : & 0 & 1 & 1 & : & 0 & 0 & 0 & : & 1 & 0 & 0 \\
g_1 & 0 & 0 & : & 0 & : & -1 & 0 & 0 & : & 0 & 0 & 0 & : & 0 & 1 & 0 \\
g_2 & 0 & 0 & : & 0 & : & 0 & 0 & -1 & : & 0 & 0 & 0 & : & 0 & 0 & 1
\end{array}$$

An inductive element L_2 is defined for edge 6

If however with the above augmentation we still cannot characterize the network in the canonical form, a knowledge of the network topology and the branch relations is not enough to determine the behavior of the network. N is hence potentially irregular and more information is needed either about the topology of N or the branch relations of the elements of N. The augmented N_M network is drawn in Fig. 5 and it is easily verified that D is indeed the complete node cutset matrix of N_M.

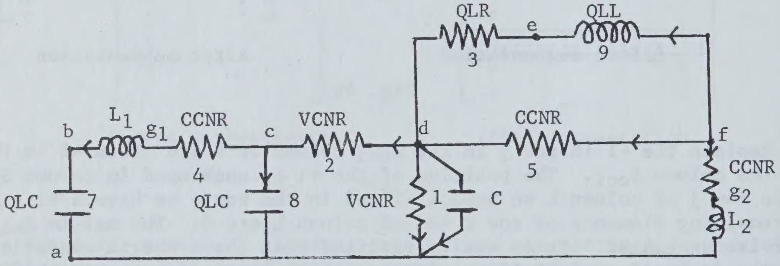


Fig. 5. The augmented network N_M

VI. CONCLUSION

Based on the work of Oh, Stern, and Meadows [5] an algorithm has been written to identify resistive irregular networks and to augment them accordingly so that the resulting modified network can be put in the canonical form:

$$\dot{\underline{x}} = \underline{f}(\underline{x},t)$$

where $\underline{x}$ is a set of properly chosen state variables.

The procedure consists essentially of manipulating matrices. These matrix operations are easily and speedily carried out on a binary computer by logical operations on packed data, since at no time is any result obtained other than 1,0,-1. As an aid to following the algorithm, an example is provided to illustrate each step of the procedure.

REFERENCES

1 Bashkow, T.R., "The A-matrix, a New Network Description", IRE Trans. on CT, Vol. CT-4, No. 3, Sept. 1957, pp. 117-119.

2 Bryant, P.R., "The Explicit Form of Bashkow's A-matrix", IRE Trans. on CT, Vol. CT-9, No. 3, Sept. 1962, pp. 303-306.

3 Stern, T.E., "On the Equations of Nonlinear Networks", IEEE Trans. on CT, Vol. CT-13, No. 1, March 1966, pp. 74-81.

4 Desoer, C.A. and Katzenelson, Jr., "Nonlinear RLC Networks", Bell System Technical Journal, Vol. XLIV, No. 1, Jan. 1965, pp. 161-198.

5 Oh, S.J., Stern, T.E., and Meadows, H.E., "On the Analysis of Nonlinear Irregular Networks", presented at the Symposium on Generalized Networks, PIB, April 1966.

6 Oh, S.J., "Analysis of Nonlinear Irregular Networks", Technical Report, No. 96, Columbia University, N.Y., June 1966.

7 Kuo, F.F. and Kaiser, J.F., Systems Analysis by Digital Computer, New York, 1967, p. 59.

A NOTE ON A CURIOUS PROPERTY OF THE EQUATIONS OF NONLINEAR NETWORKS CONTAINING TRANSISTORS

D. MITRA, I. W. SANDBERG, AND B. GOPINATH
Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

ABSTRACT

A basic property which a network with transistors, linear resistors and sources might have is that the real eigenvalues of the Jacobian matrix of the nonlinear differential equations which govern the network behavior are positive at an arbitrary operating point. The conditions under which this property holds shed considerable light on the network behavior. An interesting result which follows as a consequence of the paramountcy of the short-circuit conductance matrix of the resistive part of the network is that for networks with at most two transistors the nonnegativity of all the principal minors of a certain matrix is necessary and sufficient for all real eigenvalues of the Jacobian matrix to be positive at an arbitrary operating point; however, for three or more transistors this property need not hold and a network is provided as a counterexample.

INTRODUCTION

In the analysis of circuits containing transistors, linear resistors and sources under large signal conditions, a problem that arises in a number of different contexts concerns the behavior of the real eigenvalues of the Jacobian matrix of the nonlinear differential equations which govern the network behavior. The contexts in which the problem arises are discussed in Section 1 which also includes a theorem which relates the hypothesis of positivity of the real eigenvalues of the Jacobian matrix to conditions associated with various other problems. In Section 2 the "curious property" referred to in the title is discussed. It is shown there that as a consequence of the paramountcy of the short-circuit-conductance matrix of the resistive part of the network, networks with at most two transistors have the property that the nonnegativity of all the principal minors of a certain matrix is necessary and sufficient for the positivity of all real eigenvalues of the Jacobian matrix at an arbitrary operating point; however, for three or more transistors this property need not hold and a network is provided as a counterexample.

1. A large class of networks containing transistors, linear resistors and sources (see Fig. 1) are governed by equations[1] of the form:

$$\frac{du}{dt} + TF[C^{-1}(u)] + GC^{-1}(u) = b(t) \qquad (1)$$

where, with n the number of transistors in the network,

(i) $T = T_1 \oplus T_2 \ldots \oplus T_n$, the direct sum of n 2 by 2 matrices of the form

$$T_i = \begin{bmatrix} 1 & -\alpha_r^i \\ -\alpha_f^i & 1 \end{bmatrix}$$

with $0 \leq \alpha_r^i, \alpha_f^i < 1$ for $i = 1,2,\ldots,n$.

(ii) G, a 2n by 2n matrix, is the short circuit conductance matrix of the resistive part of the network.

(iii) $F(\cdot)$ is a mapping of E^{2n}, the set of all real 2n-vectors, into E^{2n} with the following "diagonal" structure,

$$F_i(x) = f_i(x_i), \quad i = 1,2,\ldots,2n$$

where $f_i(\cdot)$ is continuously differentiable and, furthermore, $f_i'(x) > 0$ for all $x \in E^1$. The functions $f_{2i-1}(\cdot)$ and $f_{2i}(\cdot)$ represent the static diode characteristics of the i^{th} transistor.

(iv) $C(\cdot)$ maps E^{2n} into E^{2n} and is defined thus:

$$C(x) = cx + \tau F(x) \tag{2}$$

where c and τ are diagonal matrices with positive diagonal elements. It is easily verified that the inverse $C^{-1}(\cdot)$ of $C(\cdot)$ exists.

(v) $b(\cdot)$ is a real 2n-vector-valued function which takes into account the voltage and current generators present in the network.

(vi) u is related to v the vector of transistor junction voltages through $C(v) = u$ for all $v \in E^{2n}$.

The Jacobian matrix $J(u)$ of $TF[C^{-1}(u)] + GC^{-1}(u)$ is given thus:

$$J(u) = T \operatorname{diag}\left[\frac{f_j'[C_j^{-1}(u)]}{c_j + \tau_j f_j'[C_j^{-1}(u)]}\right] + G \operatorname{diag}\left[\frac{1}{c_j + \tau_j f_j'[C_j^{-1}(u)]}\right]$$

and is of course of the form $TD_1 + GD_2$ where D_1 and D_2 are diagonal matrices with positive diagonal elements.

It is well known that the signs of the real eigenvalues of the Jacobian matrix at an equilibrium point play an important role in determining whether or not the equilibrium point is stable. Also, as shown in Ref. 2, the signs of the real eigenvalues of J(u) play a key role in the theory of the use of certain algorithms for the numerical integration of the circuit equation (1) under large step-size conditions. It has been shown[2] that in implementing the general multipoint formula for numerical integration a nonlinear equation of the following form has to be solved at each step,

$$[\tau(hb_{-1})^{-1} + T]F(x_{i+1}) + [c(hb_{-1})^{-1} + G]x_{i+1} = q_i \tag{3}$$

where h is the step size, b_{-1} is a real positive number and q_i, which includes the effect of the external source vector b(t), is considered to be an arbitrary vector in E^{2n}. Equation (3) possesses an unique solution for each q_i if and only if

$$[\tau(hb_{-1})^{-1} + T]^{-1}[c(hb_{-1})^{-1} + G] \in P_0 \tag{4}$$

where P_0 denotes the class of matrices with all principal minors nonnegative,[3] (see also Ref. 4). Now $[\tau(hb_{-1})^{-1} + T]^{-1}[c(hb_{-1})^{-1} + G]$ is of the form $(D_a+T)^{-1}(D_b+G)$ where D_a and D_b are diagonal matrices with positive diagonal elements and clearly

$$(D_a+T)^{-1}(D_b+G) \in P_0 \quad \text{for all diagonal } D_a,\ D_b > 0 \tag{5}$$

is sufficient for Eq. (3) to have an unique solution for each q_n and all step sizes. If however either τ or c or both are regarded as arbitrary then condition (5) is also necessary. The equivalence of the condition of (5) and the positivity of the real eigenvalues of J(u) at an arbitrary operating point u is established in Theorem 1 which is due to Sandberg[4] and is stated here without proof.

Theorem 1: The following statements are equivalent.

(i) The real eigenvalues of $TD_1 + GD_2$ are positive for all diagonal $D_1, D_2 > 0$.

(ii) $T^{-1}(G+D) \in P_0$ for all diagonal $D \geq 0$.

(iii) $(D_a+T)^{-1}(D_b+G) \in P_0$ for all diagonal $D_a, D_b > 0$.

(iv) $(T_w)^{-1}G_w \in P_0$ for all pairs T_w and G_w obtained from T and G, respectively, by deleting an arbitrary set w of rows and the same set of columns.

(v) $M^{-1}G \in P_0$ for all $M \in \mathcal{J}(T)$, where $\mathcal{J}(T)$ is the class of matrices obtained by replacing in T α_r^i by β_r^i and α_f^i by β_f^i where $0 \leq \beta_r^i \leq \alpha_r^i$ and $0 \leq \beta_r^i \leq \alpha_f^i$ for $i = 1,2,\ldots,n$.

(ii) and (v) are used in the following section, and statement (iv) has the following interpretation. The matrices which correspond to T and G in networks obtained by short circuiting an arbitrary number of at most all but one of the transistors' junctions are, with the appropriate choice of w, T_w and G_w respectively and $T_w^{-1}G_w \in P_0$ is both necessary and sufficient for the existence of at most one solution of the d.c. equation for the resulting network.

2. Since G is the short-circuit-conductance matrix of the linear resistive part of the network it is paramount[5*]. It is shown below that due to this property of G, for networks with at most two transistors, (i.e., for n = 1, or n = 2), it is sufficient that $T^{-1}G \in P_0$ for $T^{-1}(G+D) \in P_0$ for all diagonal $D \geq 0$, and as a consequence of Theorem 1 a number of equivalent conditions are implied. It is further shown that for networks with more than two transistors this property need not hold and a network with three transistors is given as a counter example.

The symbol $|X|_{j_1 j_2 \ldots j_p}^{i_1 i_2 \ldots i_p}$ with $1 \leq i_1 < i_2 \ldots < i_p$ and $1 \leq j_1 < j_2 \ldots < j_p$ is used to denote the minor of the matrix X corresponding to the submatrix formed by the intersection of rows $i_1, i_2, \ldots, i_p$ and columns $j_1, j_2, \ldots, j_p$.

Let n = 2 since the proof for the remaining case of n = 1 follows trivially. An obvious decomposition of D is as follows.

$$D = \sum_{i=1}^{2n} d_i E_i \tag{6}$$

where d_i (≥ 0) is the i^{th} diagonal element of D, and E_i is a matrix all elements of which are zero except the $(i,i)^{th}$ element which is 1. The principal minors of $T^{-1}(G+d_1E_1)$ may be obtained by a straightforward application of the Binet-Cauchy formula and on account of the block diagonal structure of T they have the following simple expansions:

---

[*] An n^{th} order symmetric matrix is defined to be paramount if each principal minor of order r is not less than the absolute value of any r^{th} order minor constructed from the same rows (or columns) where $r = 1,2,\ldots,n-1$.

$$|T^{-1}(G+d_1E_1)|_i^i = |T^{-1}G|_i^i + \frac{d_1\delta_{i1}}{1-\alpha_r^1\alpha_f^1} \tag{7}$$

$$|T^{-1}(G+d_1E_1)|_{ij}^{ij} = |T^{-1}G|_{ij}^{ij} + \frac{d_1\delta_{i1}}{1-\alpha_r^1\alpha_f^1}\left[|T^{-1}G|_j^j - \alpha_f^1\delta_{j2}|T^{-1}G|_j^i\right] \tag{8}$$

$$|T^{-1}(G+d_1E_1)|_{ijk}^{ijk} = |T^{-1}G|_{ijk}^{ijk} + \frac{d_1\delta_{i1}}{1-\alpha_r^1\alpha_f^1}\left[|T^{-1}G|_{jk}^{jk} - \alpha_f^1\delta_{j2}|T^{-1}G|_{jk}^{ik}\right] \tag{9}$$

where the Kronecker symbol

$$\begin{aligned}\delta_{ij} &= 0 \quad \text{if} \quad i \neq j \\ &= 1 \quad \text{if} \quad i = j.\end{aligned}$$

The remaining principal minor, namely $|T^{-1}(G+d_1E_1)|_{ijkl}^{ijkl}$ is clearly nonnegative since the determinant of T, a nonsingular matrix, is positive and $(G+d_1E_1)$ is a symmetric positive semidefinite matrix whose determinant is nonnegative. To prove $T^{-1}(G+d_1E_1) \in P_0$ it only remains to show that the bracketed terms in (8) and (9) are nonnegative. But

$$|T^{-1}G|_2^2 - \alpha_f^1|T^{-1}G|_2^1 = [G]_{22} \geq 0 \tag{10}$$

and

$$|T^{-1}G|_{2k}^{2k} - \alpha_f^1|T^{-1}G|_{2k}^{1k} = \frac{1}{1-\alpha_r^2\alpha_f^2}\left|\,|G|_{23}^{23} + \alpha_r^2|G|_{23}^{24}\right| \quad \text{for } k = 3 \tag{11}$$

$$= \frac{1}{1-\alpha_r^2\alpha_f^2}\left|\,|G|_{24}^{24} + \alpha_f^2|G|_{24}^{23}\right| \quad \text{for } k = 4 \tag{12}$$

Since G is paramount and α_r^2, $\alpha_f^2 < 1$ the expressions in (11) and (12) are nonnegative. Clearly the above arguments may be repeated to show that the paramountcy of G and $T^{-1}G \in P_0$ are sufficient for $T^{-1}(G+d_iE_i) \in P_0$ for all $d_i \geq 0$. Finally we note that $G + d_iE_i$ is also a paramount matrix and by successively adding to G the terms in the decomposition of D given in (6) the proof is completed.

The network in Fig. 2 is provided to show that the property does not hold for networks containing more than two transistors. The G matrix of this network is the following:

$$G = \begin{bmatrix} 2 & -1 & 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 & 0 \\ -1 & 1 & -1 & 2 & -1 & 0 \\ -1 & 0 & 0 & -1 & 2 & -1 \\ 1 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \tag{13}$$

Note that G is a matrix of rank 3 and, therefore it follows from an application of the Binet-Cauchy formula that all principal minors of $T^{-1}G$ of order greater than 3 have the value 0. The first order minors of $T^{-1}G$ are of the form $\frac{1}{1 - \alpha_r^i \alpha_f^i} \{[G]_{jj} - \beta\}$ where β is either α_r^i or α_f^i and are nonnegative since $[G]_{jj} \geq 1$ for $j = 1,2,\ldots,6$. It may now be verified that all principal minors of $T^{-1}G$ of orders two and three except $|T^{-1}G|_{246}^{246}$ are nonnegative and $|T^{-1}G|_{246}^{246} \geq 0$ only if

$$\frac{\alpha_f^1}{1 - \alpha_f^1} \frac{\alpha_f^3}{1 - \alpha_f^3} (1 - \alpha_f^2) \leq 1 \tag{14}$$

It is assumed that the transistors are such that the above inequality holds and, furthermore,

$$\frac{\alpha_f^1}{1 - \alpha_f^1} \frac{\alpha_f^3}{1 - \alpha_f^3} > 1 \tag{15}$$

Then there exists a β, $0 < \beta < \alpha_f^2$, such that with α_f^2 replaced by β the inequality in (14) is violated and $\hat{T}^{-1}G \notin P_0$ where $\hat{T}$ denotes the matrix obtained by replacing α_f^2 by β in T. Now $\hat{T}$ is an element of $\mathfrak{J}(T)$, the class of matrices defined in statement (v) of Theorem 1 and it follows as a consequence of that statement that the construction of the counterexample is complete.

REFERENCES

1. Sandberg, I. W., "Some Theorems on the Dynamic Response of Nonlinear Transistor Networks," B.S.T.J., 48, No. 1, January 1969, pp. 35-54.

2. Sandberg, I. W., "Theorems on the Analysis of Nonlinear Transistor Networks," B.S.T.J., 49, No. 1, January 1970, pp. 95-114.
3. Fiedler, M. and V. Pták, "Some Generalizations of Positive Definiteness and Monotonicity," Numer. Math., 9, 166, pp. 163-172.
4. Sandberg, I. W., "Theorems on the Computation of the Transient Response of Nonlinear Networks Containing Transistors and Diodes," to appear in the B.S.T.J., October 1970.
5. Cederbaum, I., "Conditions for the Impedance and Admittance Matrix of n-ports Without Ideal Transformers," Proc. IEE, pt. C, 105, March 1958, pp. 245-251.

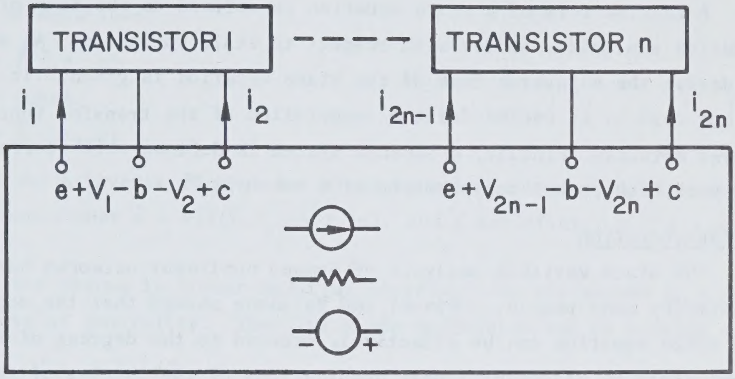

TRANSISTOR I
TRANSISTOR n
i_1
i_2
i_{2n-1}
i_{2n}
e
$+V_1$
b
$-V_2$
c
e
$+V_{2n-1}$
b
$-V_{2n}$
c
−
+

FIG. I : GENERAL NETWORK CONTAINING TRANSISTORS, RESISTORS AND SOURCES.

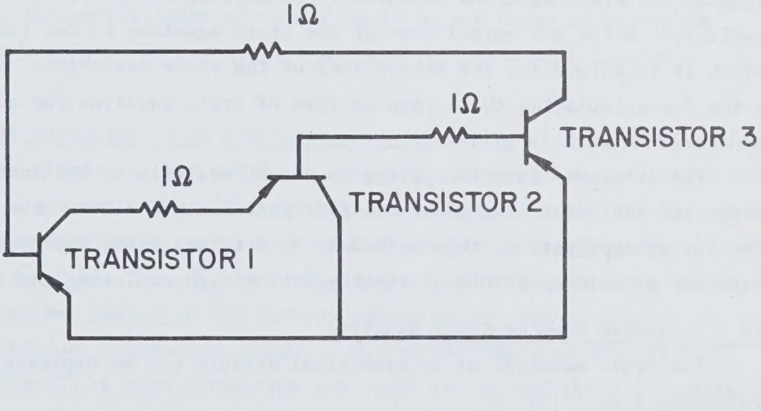

1Ω
1Ω
1Ω
TRANSISTOR 3
TRANSISTOR 2
TRANSISTOR I

FIG. 2 : THE NETWORK IN THE COUNTER EXAMPLE.

A-INVERSE FORM OF STATE EQUATION AND TRANSFER FUNCTION OF LINEAR NETWORK

YASUTOSHI ISHIZAKI
Transmission Division, Nippon Electric Company, Ltd.
Kawasaki, Japan

ABSTRACT

A-inverse form of a state equation is defined as the form of the equation explicitly solved with respect to state variables. An algorithm to derive the A-inverse form of the state equation is given. It is pointed out that it is useful for the computation of the transfer functions of linear networks. Finally, A-inverse system is defined which corresponds to low-pass high-pass transformation of a network.

I. Introduction

The state variable analysis of lumped nonlinear networks has been studied by many people. Ohtsuki and Watanabe showed that the degrees of the state equation can be effectively reduced to the degrees of freedom in Bers' sense by selecting linear combinations of the charges of the capacitive edges and linear combinations of the fluxes of the inductive edges as state variables [1]. They also presented an algorithm for calculating the state equation for the above state variables.

In this paper, the A-inverse form of state equation is defined as the form of the state equation in which it is explicitly solved for the state variables, while the normal form of the state equation is the form in which it is solved for the derivatives of the state variables. An algorithm for calculating the A-inverse form of state equation for a general nonlinear network is given.

The A-inverse form is applied to the DC analysis of nonlinear networks and the determination of transfer functions of linear networks. In the latter application, this method is used to calculate the coefficient matrices of A-inverse form of state equation without losing the sparsity.

II. A-inverse Form of State Equation

The state equation of an electrical network can be expressed in the following form : [1]

$$\begin{aligned} \dot{x} &= f(x,x_0,u) \\ \dot{x}_0 &= 0 \qquad (1) \\ y &= g(x,x_0,u) \end{aligned}$$

where x , u and y are the state variable vector, input vector and output vector, respectively. x_0 denotes state variables which do not vary with time.

Such state variables appear when the network contains cutsets composed only of capacitive edges or loops composed only of inductive edges. Since x_0 does not vary with time we can consider only x as state variables.

The form of state equation of (1) is called the normal form.

A-inverse form of a state equation is defined as follows.

Definition 1. Corresponding to the state equation (1) (in normal form), a system of equations:

$$\begin{aligned} x &= \overline{f}(\dot{x}, x_0, u) \\ x_0 &= \text{const.} \\ y &= \overline{g}(\dot{x}, x_0, u) \end{aligned} \tag{2}$$

is called the A-inverse form of the state equation, where $\overline{f}$ is an inverse of f in the sense that $\dot{x} = f(\overline{f}(\dot{x},x_0,u),x_0,u)$, and $\overline{g}$ satisfies $y = \overline{g}(f(x,x_0,u), x_0,u)$.

When the system is linear and time invariant, we can assume $x_0 = 0$ without loss of generality. Then the above definition can be written as:

$$\begin{bmatrix} \dot{x} \\ y \end{bmatrix} = \begin{bmatrix} A & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \tag{1'}$$

and

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \overline{A} & \overline{b} \\ \overline{c} & \overline{d} \end{bmatrix} \begin{bmatrix} \dot{x} \\ u \end{bmatrix} \tag{2'}$$

where $\overline{A} = A^{-1}$, $\overline{b} = -A^{-1}b$, $\overline{c} = cA^{-1}$ and $\overline{d} = d-cA^{-1}b$.

Thus, the coefficients of normal equation and those of A-inverse equation are related by a pivotal transformation with A as pivotal submatrix.

III. State Variables of An Electrical Network

Let an electrical network $\mathcal{N}$ be composed of capacitive, resistive and inductive edges and independent current and voltage sources (the dependent sources be equivalently expressed by mutual couplings between edges). We denote by $\mathcal{E}_E$, $\mathcal{E}_C$, $\mathcal{E}_R$, $\mathcal{E}_L$ and $\mathcal{E}_J$ the sets of voltage source, capacitive, resistive, inductive and current source edges, respectively.

We introduce a notation $\mathcal{N}\{\mathcal{E}_A ; \mathcal{E}_B\}$ which indicates the network obtained from $\mathcal{N}$ by open circuiting the edges in $\mathcal{E}_A$ and short circuiting the edges in $\mathcal{E}_B$, where $\mathcal{E}_A$ and $\mathcal{E}_B$ are mutually disjoint subsets of the edges of $\mathcal{N}$.

The following conditions are assumed for $\mathcal{N}$.

a) There exists in $\mathcal{N}\{\mathcal{E}_C ; \mathcal{E}_L\}$ neither a cutset which is composed only of the edges of $\mathcal{E}_J$ nor a loop which is composed only of the edges of $\mathcal{E}_E$.

b) The sum of the fluxes (charges) of each loop which consists only of inductive edges (cutset which consists only of capacitive edges) is given.

One element kind networks $\mathcal{N}_C$, $\mathcal{N}_R$, $\mathcal{N}_L$, $\overline{\mathcal{N}}_C$, $\overline{\mathcal{N}}_R$ and $\overline{\mathcal{N}}_L$ are defined as

$$\begin{aligned} \mathcal{N}_C &\triangleq \mathcal{N}\{\mathcal{E}_J \vee \mathcal{E}_L \vee \mathcal{E}_R \,;\, \mathcal{E}_E\} \\ \mathcal{N}_R &\triangleq \mathcal{N}\{\mathcal{E}_J \vee \mathcal{E}_R \,;\, \mathcal{E}_C \vee \mathcal{E}_E\} \\ \mathcal{N}_L &\triangleq \mathcal{N}\{\mathcal{E}_J \,;\, \mathcal{E}_L \vee \mathcal{E}_C \vee \mathcal{E}_E\} \end{aligned} \tag{3}$$

$$\begin{aligned} \overline{\mathcal{N}}_C &\triangleq \mathcal{N}\{\mathcal{E}_J \,;\, \mathcal{E}_R \vee \mathcal{E}_L \vee \mathcal{E}_E\} \\ \overline{\mathcal{N}}_R &\triangleq \mathcal{N}\{\mathcal{E}_J \vee \mathcal{E}_C \,;\, \mathcal{E}_L \vee \mathcal{E}_E\} \\ \overline{\mathcal{N}}_L &\triangleq \mathcal{N}\{\mathcal{E}_J \vee \mathcal{E}_C \vee \mathcal{E}_R \,;\, \mathcal{E}_E\} \end{aligned} \tag{4}$$

Let T_C be a C-normal tree of $\mathcal{N}$ and T_L be an L-normal tree such that 1) all inductive branches of T_C are branches of T_L and 2) all capacitive chords of T_C are chords of T_L [1].

The sets of edges $\mathcal{E}_C$, $\mathcal{E}_L$ and $\mathcal{E}_R$ are partitioned by means of T_C and T_L as follows :

$$\mathcal{E}_R^{\epsilon} \in T_L , \qquad \mathcal{E}_R^{\zeta} \in \widetilde{T}_L \tag{5}$$

$$\begin{aligned} &\mathcal{E}_C^{\xi} , \mathcal{E}_L^{\xi} \in T_C \cap T_L , \qquad \mathcal{E}_C^{\zeta} , \mathcal{E}_L^{\zeta} \in T_C \cap \widetilde{T}_L \\ &\mathcal{E}_C^{\epsilon} , \mathcal{E}_L^{\epsilon} \in \widetilde{T}_C \cap T_L , \qquad \mathcal{E}_C^{\eta} , \mathcal{E}_L^{\eta} \in \widetilde{T}_C \cap \widetilde{T}_L \end{aligned} \tag{6}$$

where $\widetilde{T}_C$ and $\widetilde{T}_L$ are the cotrees of T_C and T_L, respectively. Then, from the above definition of T_C and T_L, $\mathcal{E}_L^{\zeta}$ and $\mathcal{E}_C^{\epsilon}$ are empty. The Kirchhoff's laws based on T_L can be written as :

$$\begin{bmatrix} I_C^{\xi} \\ I_R^{\epsilon} \\ I_L^{\xi} \\ I_L^{\epsilon} \\ I_E \end{bmatrix} = - \begin{bmatrix} 0 & 0 & \overline{N}_{13} & \overline{N}_{14} & 0 \\ 0 & \overline{N}_{22} & \overline{N}_{23} & \overline{N}_{24} & \overline{N}_{25} \\ \overline{N}_{31} & \overline{N}_{32} & \overline{N}_{33} & \overline{N}_{34} & \overline{N}_{35} \\ \overline{N}_{41} & \overline{N}_{42} & \overline{N}_{43} & \overline{N}_{44} & \overline{N}_{45} \\ 0 & \overline{N}_{52} & \overline{N}_{53} & \overline{N}_{54} & \overline{N}_{55} \end{bmatrix} \begin{bmatrix} I_L^{\eta} \\ I_R^{\zeta} \\ I_C^{\eta} \\ I_C^{\zeta} \\ J \end{bmatrix}. \tag{7}$$

$$\begin{bmatrix} V_L^{\eta} \\ V_R^{\zeta} \\ V_C^{\eta} \\ V_C^{\zeta} \\ V_J \end{bmatrix} = + \begin{bmatrix} 0 & 0 & \overline{N}'_{31} & \overline{N}'_{41} & 0 \\ 0 & \overline{N}'_{22} & \overline{N}'_{32} & \overline{N}'_{42} & \overline{N}'_{52} \\ \overline{N}'_{13} & \overline{N}'_{23} & \overline{N}'_{33} & \overline{N}'_{43} & \overline{N}'_{53} \\ \overline{N}'_{14} & \overline{N}'_{24} & \overline{N}'_{34} & \overline{N}'_{44} & \overline{N}'_{54} \\ 0 & \overline{N}'_{25} & \overline{N}'_{35} & \overline{N}'_{45} & \overline{N}'_{55} \end{bmatrix} \begin{bmatrix} V_C^{\xi} \\ V_R^{\epsilon} \\ V_L^{\xi} \\ V_L^{\epsilon} \\ E \end{bmatrix} \tag{8}$$

In the above coefficient matrices, the zeros at (1,1) (1,2) and (2,1) entries are due to the definitions of T_L and T_C and the zeros at (1,5) and (5,1) entries are the results of excluding the J-only cutset and the E-only loop in $\mathcal{N}\{\mathcal{E}_C ; \mathcal{E}_L\}$. (Condition (a))

On the other hand, based on T_C, the currents of the edges in $\mathcal{E}_L^{\xi}$ and the voltages of the edges in $\mathcal{E}_C^{\eta}$ can be expressed as:

$$I_L^{\xi} = -N_{13}I_L^{\eta} - N_{14}I_L^{\epsilon} - N_{15}J \qquad (9)$$
$$V_C^{\eta} = N'_{31}V_C^{\xi} + N'_{41}V_C^{\zeta} + N'_{51}E \qquad (10)$$

From (7) and (10), utilizing the property that a cutset and a loop are orthogonal each other, we have

$$\bar{N}_{23} - \bar{N}_{24}\, N_{41} = 0 \qquad (11)$$
$$\begin{bmatrix}\bar{N}_{33}\\ \bar{N}_{43}\end{bmatrix} - \begin{bmatrix}\bar{N}_{34}\\ \bar{N}_{44}\end{bmatrix} N_{41} = 0 \qquad (12)$$

Similarly, from (8) and (9),

$$\bar{N}'_{32} + \bar{N}'_{42}\, N'_{14} = 0 \qquad (13)$$
$$\begin{bmatrix}\bar{N}'_{33}\\ \bar{N}'_{34}\end{bmatrix} + \begin{bmatrix}\bar{N}'_{43}\\ \bar{N}'_{44}\end{bmatrix} N'_{14} = 0 \qquad (14)$$

As shown in reference [1], the sum of inductive fluxes in each loop corresponding to $\mathcal{E}_L^{\epsilon}$ and the sum of capacitive charges in each cutset corresponding to $\mathcal{E}_C^{\xi}$ can be taken as state variables :

$$x \triangleq \begin{bmatrix}\Phi^*\\ Q^*\end{bmatrix} \triangleq \begin{bmatrix}\Phi_L^{\epsilon} - N'_{14}\Phi_L^{\xi}\\ Q_C^{\zeta} + N_{41}\, Q_C^{\eta}\end{bmatrix} \qquad (15)$$

The given charges and fluxes in the condition b) are denoted Q_o and Φ_o, respectively, and collectively denoted by x_o.

$$x_o \triangleq \begin{bmatrix}\Phi_o\\ Q_o\end{bmatrix} \qquad (16)$$

The currents J of the current source edges and the voltages E of the voltage source edges are considered to be the input to $\mathcal{N}$, and denoted by u.

$$u \triangleq \begin{bmatrix}E\\ J\end{bmatrix} \qquad (17)$$

In reference [1], an algorithm for calculating $\dot{x}$ from given x, x_o and u is presented, which is equivalent to give the function f. The algorithm consists of analyses of $\mathcal{N}_C$ and $\mathcal{N}_L$ at the first stage and $\mathcal{N}_R$ at the next.

In the following, we will show an algorithm for calculating x from given $\dot{x}$, x_o and u, which is equivalent to give the A-inverse form of the state equation, $\bar{f}$.

IV. Derivation of A-inverse Form of State Equation

By the assumption

$$\dot{\Phi}^* = V_L^{\epsilon} - N'_{14}\, V_L^{\xi} \qquad (18)$$
$$\dot{Q}^* = I_C^{\zeta} + N_{41}\, I_C^{\eta} \qquad (19)$$

are given.

The algorithm consists of analyses of three one-element-kind networks $\overline{\mathcal{N}}_R$, $\overline{\mathcal{N}}_C$ and $\overline{\mathcal{N}}_L$.

Analysis of $\overline{\mathcal{N}}_R$

The second equation of (7) and the second equation of (8) can be rewritten as :

$$I_R^{\epsilon} + \overline{N}_{22} I_R^{\zeta} = -\overline{N}_{24}(N_{41} I_C^{\eta} + I_C^{\zeta}) - \overline{N}_{25} J = -\overline{N}_{24}\dot{Q}^* - \overline{N}_{25} J \tag{20}$$

$$V_R^{\zeta} - \overline{N}'_{22} V_R^{\epsilon} = \overline{N}'_{42}(-N'_{14} V_L^{\xi} + V_L^{\epsilon}) + \overline{N}'_{52} E = \overline{N}'_{42}\dot{\Phi}^* + \overline{N}'_{52} E \tag{21}$$

For a given set of $\dot{Q}^*$, $\dot{\Phi}^*$, J and E, equations (20) and (21) give the Kirchhoff's current law and voltage law, respectively, of $\overline{\mathcal{N}}_R$, where $-N_{24}\dot{Q}^*-N_{25}J$ and $N'_{42}\dot{\Phi}^*+N'_{52}E$ are regarded as the equivalent current sources inserted in parallel with the edges of $\mathcal{E}_R^{\epsilon}$ and the equivalent voltage sources inserted in series with the edges of $\mathcal{E}_R^{\zeta}$, respectively.

Let the characteristics of the edges of $\mathcal{E}_R$ be given by

$$\begin{bmatrix} I_R^1 \\ V_R^2 \end{bmatrix} = H_R(\begin{bmatrix} V_R^1 \\ I_R^2 \end{bmatrix}) \tag{22}$$

where I_R^1 and V_R^1 are the currents and voltages of some edges of $\mathcal{E}_R$, respectively, and I_R^2 and V_R^2 are the currents and the voltages of the other edges of $\mathcal{E}_R$, respectively.

When the condition for a unique solution of one-element-kind network [1] is satisfied for $\overline{\mathcal{N}}_R$, I_R^{ϵ}, I_R^{ζ}, V_R^{ϵ} and V_R^{ζ} can be expressed as functions of $\dot{Q}^*$, $\dot{\Phi}^*$, J and E by solving (20), (21) and (22).

Analysis of $\overline{\mathcal{N}}_C$

From the first equation of (7) and the third and forth equations of (8),

$$I_C^{\xi} + \overline{N}_{13} I_C^{\eta} + \overline{N}_{14} I_C^{\zeta} = 0 \tag{23}$$

$$\begin{bmatrix} V_C^{\eta} \\ V_C^{\zeta} \end{bmatrix} - \begin{bmatrix} \overline{N}'_{13} \\ \overline{N}'_{14} \end{bmatrix} V_C^{\xi} = \begin{bmatrix} \overline{N}'_{23} \\ \overline{N}'_{24} \end{bmatrix} V_R^{\epsilon} + \begin{bmatrix} \overline{N}'_{33} \\ \overline{N}'_{34} \end{bmatrix} V_L^{\xi} + \begin{bmatrix} \overline{N}'_{43} \\ \overline{N}'_{44} \end{bmatrix} V_L^{\epsilon} + \begin{bmatrix} \overline{N}'_{53} \\ \overline{N}'_{54} \end{bmatrix} E \tag{24}$$

Integrating (23) with respect to time, we have

$$Q_C^{\xi} + \overline{N}_{13} Q_C^{\eta} + \overline{N}_{14} Q_C^{\zeta} = Q_o \tag{25}$$

where Q_o is an integration constant which is given by condition b)

Putting (12) into (24) and taking into account that $\dot{\Phi}_L^{\xi} = V_L^{\xi}$, $\dot{\Phi}_L^{\epsilon} = V_L^{\epsilon}$, and (15), we have

$$\begin{bmatrix} V_C^{\eta} \\ V_C^{\zeta} \end{bmatrix} - \begin{bmatrix} \overline{N}'_{13} \\ \overline{N}'_{14} \end{bmatrix} V_C^{\xi} = \begin{bmatrix} \overline{N}'_{23} \\ \overline{N}'_{24} \end{bmatrix} V_R^{\epsilon} + \begin{bmatrix} \overline{N}'_{43} \\ \overline{N}'_{44} \end{bmatrix} \dot{Q}^* + \begin{bmatrix} \overline{N}'_{53} \\ \overline{N}'_{54} \end{bmatrix} E \tag{26}$$

Let the characteristics of the edges in $\mathcal{E}_C$ be expressed as

$$\begin{bmatrix} Q_C^1 \\ V_C^2 \end{bmatrix} = H_C(\begin{bmatrix} V_C^1 \\ Q_C^2 \end{bmatrix}) \tag{27}$$

where Q_C^1 and V_C^1 are the charges and the voltages of some edges of $\mathcal{E}_C$, respectively, and Q_C^2 and V_C^2 are those of the other edges of $\mathcal{E}_C$, respectively.

For a given set of $\dot{\Phi}^*$, Q_o and E and the solution V_R^ξ of $\bar{\mathcal{N}}_R$, equations (25) and (26) give Kirchhoff's laws for $\bar{\mathcal{N}}_C$, where Q_o and $\begin{bmatrix} \bar{N}'_{23} \\ \bar{N}'_{24} \end{bmatrix} V_R^\epsilon$ + $\begin{bmatrix} \bar{N}'_{43} \\ \bar{N}_{44} \end{bmatrix} \dot{\Phi}^*$ + $\begin{bmatrix} N'_{53} \\ N_{54} \end{bmatrix}$ E are regarded as equivalent current sources inserted in parallel with the edges of $\mathcal{E}_C^\xi$ and equivalent voltage sources inserted in series with the edges of $\mathcal{E}_C^\eta \cup \mathcal{E}_C^\zeta$, respectively.

Again assuming the condition for a unique solution for $\bar{\mathcal{N}}_C$, Q_C can be expressed as a function of Q_o, E, $\dot{\Phi}^*$ and V_R^ϵ, and hence as a function of Q_o, J, E, $\dot{\Phi}^*$ and $\dot{Q}^*$.

Analysis of $\bar{\mathcal{N}}_L$

Analysis of $\bar{\mathcal{N}}_L$ is quite similar to that of $\bar{\mathcal{N}}_C$. As a result, Φ^* can be expressed as a function of Φ_o, J, E, $\dot{Q}^*$ and $\dot{\Phi}^*$.

We have shown that for a general nonlinear lumped network, the state variables x can be calculated from their derivatives $\dot{x}$, inputs u and constants x_o by analyzing three one-element-kind networks $\bar{\mathcal{N}}_R$, $\bar{\mathcal{N}}_C$ and $\bar{\mathcal{N}}_L$.

The flow of the algorithm is shown in Fig. 1 together with the flow of calculating $\dot{x}$ from x, u and x_o.

V. DC Analysis of Nonlinear Networks

DC analysis of a nonlinear network is nothing but to determine the currents, voltages, charges and fluxes which make the derivatives of state variables be zero.

From the arguments in the preceding section, these variables can be determined by analyzing $\bar{\mathcal{N}}_R$, $\bar{\mathcal{N}}_C$ and $\bar{\mathcal{N}}_L$ for $\dot{x}=0$. As an obvious consequence, the following theorem is valid.

Theorem 1. A sufficient condition for the network $\mathcal{N}$ to have a unique DC solution is that the one-element-kind networks $\bar{\mathcal{N}}_R$, $\bar{\mathcal{N}}_C$ and $\bar{\mathcal{N}}_L$ have unique solutions.

VI. Computation of Transfer Function

In this section, we restrict ourselves to linear time invariant networks. Then, the normal form and A-inverse form of the state equation are given by (1') and (2'), respectively.

The coefficient matrices of A-inverse form of the state equation, $\bar{A}$, $\bar{b}$, $\bar{c}$ and $\bar{d}$, can be calculated by the algorithm in section III without executing the pivotal transformation. The sparsity of the coefficient matrices of A-inverse form, which is important in the following arguments, depends on the topological connection and the mutual couplings of the network. In some cases, the coefficient matrix of the inverse form may be more sparse than that of normal form. The technique described preserves the sparsity of the matrix of A-inverse form, which would be lost if pivotal transformation were used because of rounding errors in floating point arithmetic.

The transfer function F(s) = y/u of (1') can be written as

$$F(s) = K \frac{s^{\beta} \prod_{i=1}^{m} (s - s_i')}{\prod_{i=1}^{n} (s - s_i)} \tag{28}$$

where n is the order of the state equation or the degrees of freedom of the network, $m=n-\alpha-\beta$, and α and β are the orders of transmission zeros at the infinity and at the origin, respectively. β can be determined by the following theorem.

Theorem 2. If $\bar{d} \neq 0$, then $\beta = 0$.
If $\bar{d} = 0$,
$\bar{c}\bar{A}^{j-1}\bar{b} = 0$ for $1 \leq j < j_0$ and
$\bar{c}\bar{A}^{j_0-1}\bar{b} \neq 0$,
then $\beta = j_0$.

The β defined above can be considered to be consists of two parts, β_0 and β_1. β_0 is defined by $\beta_0 = \min \beta$, where minimum is taken over all the networks which have the same topology and same locations of mutual couplings as the network considered. Then β_0 is determined only from the sparsity of $\bar{A}$, $\bar{b}$, $\bar{c}$ and $\bar{d}$, which is preserved if they are calculated by the method of Section III.

β_1 represents the order of transmission zero at the origin which appears when there exists a balancing of network element values. Such a balancing may not hold exactly in a physical network.

In the method in reference [2], the roots of the numerator are obtained by solving an eigenvalue problem of a matrix of order $n - \alpha$. By the following thorem, however, we can reduce it to $n - \alpha - \beta$.

Theorem 3. If the system described by the state equation (1') has the transfer function F(s), the system:

$$\begin{aligned} \dot{\xi} &= A\xi + b\omega \\ \eta &= cA^{-\beta}\xi \end{aligned} \tag{29}$$

has the transfer function $F^*(s) = \eta/\omega = s^{-\beta}F(s)$.

Again, $cA^{-\beta} = \bar{c}\bar{A}^{\beta-1}$ can be calculated from the coefficient matrices of the A-inverse form.

VII. The A-inverse system

In this section, we introduce the concept of A-inverse system.

Definition 2. For the system given by (1'),

$$\begin{bmatrix} \dot{\bar{x}} \\ \bar{y} \end{bmatrix} = \begin{bmatrix} \bar{A} & \bar{b} \\ \bar{c} & \bar{d} \end{bmatrix} \begin{bmatrix} \bar{x} \\ \bar{u} \end{bmatrix} \tag{30}$$

is called A-inverse system of (1')

Let the original system and the A-inverse system be denoted by S and $\bar{S}$, respectively, then the following theorem is obvious consequence of the definition.

Theorem 4. $S = \bar{\bar{S}}$

Let $F(s) = y/u$ and $\bar{F}(s) = \bar{y}/\bar{u}$ be the transfer functions of the original system and the A-inverse system, respectively, then, we have:

Theorem 5. $\bar{F}(s) = F(s^{-1})$ (31)

In the following, it is shown that when the state equation (1') represents a linear time invariant network, the equation (30) becomes the state equation of the network derived from the original network by changing capacitances to inductances and inductances to capacitances by properly selecting the state variables.

Let the edge characteristics of the original network $\mathcal{N}$ be expressed in hybrid forms :

$$\begin{bmatrix} Q_C^1 \\ V_C^2 \end{bmatrix} = H_C \begin{bmatrix} V_C^1 \\ Q_C^2 \end{bmatrix} \qquad \text{(capacitive edges)} \tag{32}$$

$$\begin{bmatrix} I_R^1 \\ V_R^2 \end{bmatrix} = H_R \begin{bmatrix} V_R^1 \\ I_R^2 \end{bmatrix} \qquad \text{(resistive edges)} \tag{33}$$

$$\begin{bmatrix} I_L^1 \\ \phi_L^2 \end{bmatrix} = H_L \begin{bmatrix} \phi_L^1 \\ I_L^2 \end{bmatrix} \qquad \text{(inductive edges)} \tag{34}$$

Here, we introduce the concept of A-inverse network.

Definition 3.

The network $\overline{\mathcal{N}}$ which satisfies the following conditions is called the A-inverse network of $\mathcal{N}$.

1) $\overline{\mathcal{N}}$ has the same topological connection as $\mathcal{N}$.

2) The capacitive, resistive and inductive edges of $\mathcal{N}$ are replaced by inductive, resistive and capacitive edges in $\overline{\mathcal{N}}$, respectively, while the voltage and current source edges of $\mathcal{N}$ remain the voltage and current source edges in $\overline{\mathcal{N}}$, respectively.

3) The edge characteristics of $\overline{\mathcal{N}}$ are given by

$$\begin{bmatrix}\overline{I}_L^1\\ \overline{\Phi}_L^2\end{bmatrix} = \overline{H}_L \begin{bmatrix}\overline{\Phi}_L^1\\ \overline{I}_L^2\end{bmatrix} \qquad \text{(inductive edges)} \qquad (35)$$

$$\begin{bmatrix}\overline{I}_R^1\\ \overline{V}_R^2\end{bmatrix} = \overline{H}_R \begin{bmatrix}\overline{V}_R^1\\ \overline{I}_R^2\end{bmatrix} \qquad \text{(resistive edges)} \qquad (36)$$

$$\begin{bmatrix}\overline{Q}_C^1\\ \overline{V}_C^2\end{bmatrix} = \overline{H}_C \begin{bmatrix}\overline{V}_C^1\\ \overline{Q}_C^2\end{bmatrix} \qquad \text{(capacitive edges)} \qquad (37)$$

with $\overline{H}_L = H_C$, $\overline{H}_R = H_R$ and $\overline{H}_C = H_L$.

Let x be a state variable of $\mathcal{N}$ which consists of the sum of fluxes in each loop corresponding to $\mathcal{E}_L^{\epsilon}$ and the sum of charges in each cutset corresponding to $\mathcal{E}_C^{\xi}$ as defined by equation (15) and let $\overline{x}$ be a vector which consists of the sum of the voltages of the capacitive edges in each loop of $\overline{\mathcal{N}}$ corresponding to $\mathcal{E}_L^{\epsilon}$ ($\mathcal{E}_L^{\epsilon}$ consists of capacitive edges in $\overline{\mathcal{N}}$) and the sum of the currents of the inductive edges in each cutset of $\overline{\mathcal{N}}$ corresponding to $\mathcal{E}_C^{\xi}$ ($\mathcal{E}_C^{\xi}$ consists of inductive edges in $\overline{\mathcal{N}}$), then the following theorem holds.

Theorem 6. $\overline{x}$ is a state variable vector of $\overline{\mathcal{N}}$, and the state equation of $\overline{\mathcal{N}}$ with respect to $\overline{x}$ becomes the A-inverse of the state equation of $\mathcal{N}$ with respect to x.

From the theorems 5 and 6, it is seen that the concept of the A-inverse system corresponds to the so-called " lowpass-highpass transformation " of the linear networks.

Acknowledgment

The author wishes to express his thanks to his colleagues, Mr. Tatsuo Ohtsuki and Mr. Kenji Kawakita for their suggestions and discussions.

References

[1] T.Ohtsuki and H. Watanabe, "State variable analysis of RLC networks containing nonlinear coupling elements", IEEE Trans. on Circuit Theory vol. CT-16 no.1, pp 26-38.

[2] I.W. Sandberg and H. C. So, "A two-sets-of-eigenvalues approach to the computer analysis of linear systems", IEEE Trans. on Circuit Theory vol. CT-16 no.4, pp 509-517.

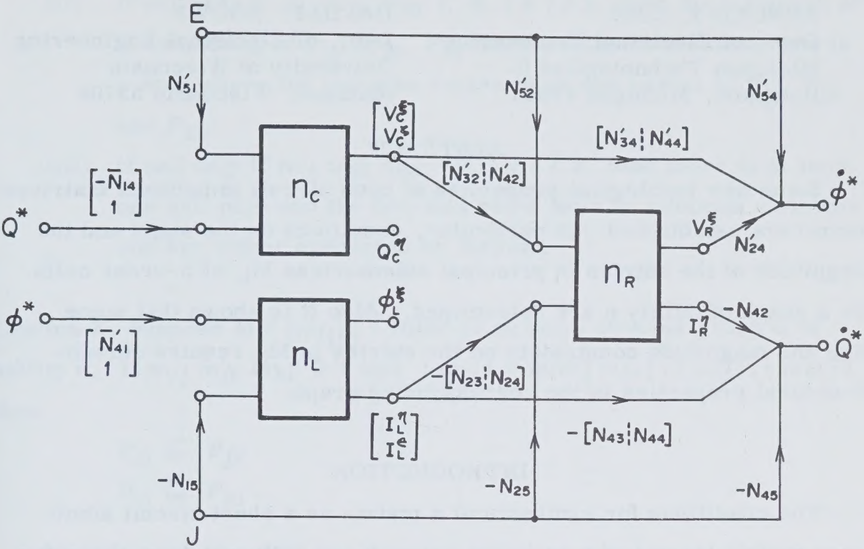

E
N'51
N'52
N'54
[V_C^ξ V_C^ξ]
[N'34 ¦ N'44]
[N'32 ¦ N'42]
[-N̄14 1]
Q*
n_C
Q_C^η
V_R^ξ
N'24
φ̇*
n_R
φ_L^ξ
φ*
[N̄'41 1]
n_L
-N42
I_R^η
Q̇*
-[N23 ¦ N24]
[I_L^η I_L^e]
-[N43 ¦ N44]
-N15
-N25
-N45
J

Fig.1 (a) Flow of the computation of the normal form of state equation.

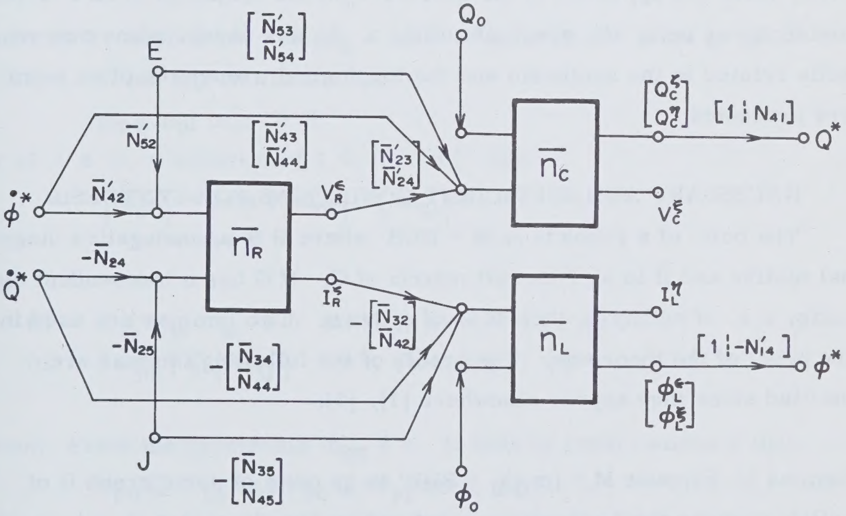

[N̄'53 N̄'54]
Q0
E
N̄'52
[N̄'43 N̄'44]
[N̄'23 N̄'24]
[Q_C^ξ Q_C^η]
[1 ¦ N41]
Q*
n̄_C
V_C^ξ
φ̇*
N̄'42
n̄_R
V_R^ε
Q̇*
-N̄24
I_R^ξ
-[N̄32 N̄42]
I_L^η
-N̄25
-[N̄34 N̄44]
n̄_L
[1 ¦ -N'14]
φ*
[φ_L^ε φ_L^ξ]
J
-[N̄35 N̄45]
φ0

Fig. 1 (b) Flow of the computation of the A-inverse form of state equation.

SOME NEW TOPOLOGICAL PROPERTIES OF OPEN CIRCUIT IMPEDANCE MATRICES

ARNOLD Y. LEE
Dept. of Electrical Engineering
Michigan Technological U.
Houghton, Michigan 49931

DAVID P. BROWN
Dept. of Electrical Engineering
University of Wisconsin
Madison, Wisconsin 53706

ABSTRACT

Some new topological properties of open circuit impedance matrices (ocim) are established. In particular, conditions on the signs and the magnitude of the entries in principal submatrices M_{ii} of n-order ocim for a graph of nullity n are determined. Also it is shown that some sign and magnitude constraints on the entries in M_{ii} require certain structural properties in the corresponding graph.

INTRODUCTION

The conditions for synthesis of a matrix as a short circuit admittance matrix (scam) of a resistive network are different from those for an open circuit impedance matrix (ocim). The conditions for synthesis of an n-order matrix as a scam of a graph of rank n are known. However, little has appeared in the literature on the synthesis of an n-order matrix as an ocim of a graph of nullity n. In this paper, some new results related to the synthesis and the topological properties of an ocim are presented.

NECESSARY AND SUFFICIENT CONDITIONS FOR SYNTHESIS

The ocim of a graph G is M = BRB' where R is a nonnegative diagonal matrix and B is an f-circuit matrix of G. If G has n independent circuits, i. e. of nullity n, then M is of order n. Two lemmas are used in the proof of the theorems. The proofs of the following lemmas are omitted since they appear elsewhere [1], [2].

Lemma 1: Suppose $M = [m_{ij}]_n = BRB'$ is an ocim of some graph G of nullity n where the f-circuits are defined by chords c_i, $i=1,2,\ldots,n$, of G. Let P_{ij} be the common elements between f-circuits defined by c_i and c_j; $i,j = 1,2,\ldots,n$, $i \neq j$.

(i) If and only if $m_{ij}\, m_{jk}\, m_{ki} > 0$, $i \neq j \neq k$, then the subgraph of G consisting of P_{ij}, P_{jk} and P_{ki} is a path.

(ii) If and only if $m_{ij}\, m_{jk}\, m_{ki} < 0$, $i \neq j \neq k$, then the subgraph of G consisting of P_{ij}, P_{jk}, P_{ki} has no common elements and is a star where the common vertex is an end vertex of P_{ij}, P_{jk} and P_{ki}.

(iii) If and only if $m_{ij}\, m_{jk}\, m_{ki} = 0$, $i \neq j \neq k$, then there is at least one null path and the non-null paths have no common elements and are either connected or disjoint.

Lemma 2: Suppose $M = [m_{ij}]_n = BRB'$ is an ocim of some graph G of nullity n. If $m_{ij}\, m_{jk}\, m_{ki} > 0$ and $|m_{ij}| = \min(|m_{jk}|, |m_{ij}|, |m_{ki}|)$, then

$$P_{ij} \subseteq P_{jk}$$
$$P_{ij} \subseteq P_{ki} .$$

Theorems 1, 2, and 3 give the constraints on the signs and magnitude of the entries of a principal submatrix of M determined from the signs and magnitude of the entries in only two rows of M.

Theorem 1: Suppose $M = [m_{ij}]_n = BRB'$ is an ocim for some graph of nullity n. If for any two indices p, q,

$$m_{pq}\, m_{pl}\, m_{ql} > 0$$

for all $l \in L$, a subset of $\{1, 2, \ldots, n\}$ and

$$|m_{pq}| = \min_{l \in L} (|m_{pl}|, |m_{ql}|)$$

then

(i) $m_{ij}\, m_{jk}\, m_{ki} > 0$ for $i, j, k \in L + \{p, q\}$

(ii) $|m_{pq}| \leqslant |m_{ij}|$ for $i, j \in L + \{p, q\}$

Proof: From the hypothesis $m_{pq} \neq 0$. It follows from Lemma 2 that

$$P_{pq} \subseteq P_{pi} \text{ and } P_{pq} \subseteq P_{pj} \text{ for } i, j \in L. \tag{1}$$

This implies that there are common elements between the paths P_{pi} and P_{pj}, hence from Lemma 1

$$m_{pi}\, m_{pj}\, m_{ij} > 0 \text{ for all } i,\ j \notin L$$

A similar argument establishes that

$$m_{qi}\, m_{qj}\, m_{ij} > 0 \text{ for all } i,\ j \in L$$

Since $m_{pi}\, m_{pj}\, m_{ij} > 0$, there are common elements P among P_{pi}, P_{pj} and P_{ij}. Two cases are considered:

Case I: Suppose P is P_{pi} or P_{pj}. It follows that P_{pi} (or P_{pj}) $\subseteq P_{ij}$. Combining (1) and (2) results in

$$P_{pq} \subseteq P_{ij}.$$

Case II: Suppose P is P_{ij}. The elements of P_{ij} are all common elements between P_{pi} and P_{pj}. From (1) some of these common elements are in P_{pq}, therefore

$$P_{pq} \subseteq P_{ij}.$$

Hence

$|m_{pq}| \leqslant |m_{ij}|$ for all $i,\ j \in L + \{p,\ q\}$ and conclusion (ii) is established. The elements of P_{pq} are common elements among the paths P_{ij}, $i,\ j \in L + \{p,\ q\}$, .

$m_{ij}\, m_{jk}\, m_{ki} \nleq 0$ for all $i,\ j \in L + \{p,\ q\}$ and conclusion (i) follows immediately.

Corollary 1: Suppose a matrix with positive entries

$M = [m_{ij}]_n = BRB'$ is an ocim of a graph of nullity n.

If $m_{pq} = \min\limits_{\forall j} (m_{pj},\ m_{qj})$

$m_{rs} = \min\limits_{\forall j} (m_{rj},\ m_{sj}) \qquad s,\ r \neq p \text{ or } q$

then

$$m_{pq} = m_{rs}.$$

Example 1: The matrix given in (3) is not an ocim of a graph of nullity since m_{14} is the minimum entry among the entries of rows 1 and 4, but m_{14} is not a minimum of M.

$$M = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 5 & 0 & 2 \\ 2 & 0 & 5 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix} \tag{3}$$

Theorem 2: Suppose $M = [m_{ij}]_n = BRB'$ is an ocim for some graph of nullity n. If for any two indices p, q

$$m_{pq}\, m_{pl}\, m_{ql} < 0$$

for all $l \in L$, a subset of $\{1, 2, \ldots, n\}$, then for $i, j \in L$ either

(i) $m_{si}\, m_{sj}\, m_{ij} > 0$ for $s = p, q$

or (ii) $m_{ij} = 0$

Proof: Without loss of generality, let $s = p$. For $i, j \in L$ from the hypothesis, P_{pi} and P_{pj} are not null paths. They either have common elements or are vertex disjoint paths otherwise circuit p has a vertex of degree three which is impossible. Hence

$$m_{pi}\, m_{pj}\, m_{ij} \nless 0$$

and the conclusions follow.

Theorem 3: Suppose $M = [m_{ij}]_n = BRB'$ is an ocim for some graph of nullity n. If for any two indices p, q

$$m_{pq}\, m_{pl}\, m_{ql} < 0$$

for all $l \in L$, a subset of $\{1, 2, \ldots, n\}$ then for $i, j, k \in L$ either

(i) $m_{ij}\, m_{jk}\, m_{ki} > 0$

or (ii) two of the three quantities in (i) are zero and the third is not zero.

Proof: Suppose for $i, j, k \in L$ all three quantities m_{ij}, m_{jk}, m_{ki} are nonzero. From Theorem 2,

$$m_{pi}\, m_{pj}\, m_{ij}\, m_{pj}\, m_{pk}\, m_{jk}\, m_{pi}\, m_{pk}\, m_{ik} > 0$$

or

$$m_{ij}\, m_{jk}\, m_{ki} > 0.$$

Suppose $m_{ij} = 0$, from the hypothesis, circuit k has common elements with either circuit i or circuit j but not both. Hence either

$$m_{ik} = 0 \qquad m_{jk} \neq 0$$

or $$m_{jk} = 0 \qquad m_{ik} \neq 0$$

and the conclusion follows.

Example 2: The matrix M given in (4) is not an ocim of a graph of nullity 4 since it contradicts Theorem 2. The same conclusion can be obtained from the fact that M is a short circuit admittance matrix of a nonplanar graph of rank 4.

$$M = \begin{bmatrix} 15 & -3 & -2 & -6 \\ -3 & 14 & -4 & -5 \\ -2 & -4 & -9 & -1 \\ -6 & -5 & -1 & 13 \end{bmatrix} \tag{4}$$

The necessary and sufficient conditions for synthesis of a matrix as an ocim of a planar graph with a specified cotree structure is next established. The conditions are a restricted form of hyperdominantcy and this result corrects a theorem appearing in the technique literature [3]. Before stating the theorem, two classes of hyperdominant matrices are defined next.

Definition 1: A hyperdominant matrix $M = [m_{ij}]_n$ which satisfies (1) and (2) below is called a $K_{3,3}$ matrix.

(1) There exists a sequence of indexes
$(l_1, l_2, \ldots, l_m)$, $m \geqslant 5$ such that
$m_{l_1 l_2}, m_{l_2 l_3}, m_{l_3 l_4}, \ldots, m_{l_{m-1} l_m}$ are nonzero.

(2) There exists $i < j < t < k < l$ and two other sequences of indexes $(l_i, p_1, p_2, \ldots, p_r, l_k)$ and $(l_j, q_1, q_2, \ldots, q_s, l_l)$, where all l's, p's, and q's are distinct, such that
$m_{l_i p_1}, m_{p_1 p_2}, m_{p_2 p_3}, \ldots, m_{p_r l_k}$ are nonzero
and
$m_{l_j q_1}, m_{q_1 q_2}, \ldots, m_{q_s l_l}$ are nonzero.

Definition 2: A hyperdominant matrix $M = [m_{ij}]_n$ which satisfies (1) and (2) below is called a K_5 matrix.

(1) There exists a sequence of indexes

$(l_1, l_2, \ldots, l_m)$ $m \geqslant 4$ such that

$m_{l_1 l_2}, m_{l_2 l_3}, \ldots, m_{l_{m-1} l_m}$ are nonzero.

(2) There exists $i < j < k < l$ and three other sequences

$(l_i, p_1, p_2, \ldots, p_r, l_k)$, $(l_j, q_1, q_2, \ldots, q_s, l_1)$

and $(l_1, t_1, t_2, \ldots, t_u, l_i)$ where all the l's, p's, q's and t's are distinct, such that

$m_{l_i p_1}, m_{p_1 p_2}, \ldots, m_{p_r l_k}$ are nonzero

$m_{l_j q_1}, m_{q_1 q_2}, \ldots, m_{q_s l_1}$ are nonzero

$m_{l_1 t_1}, m_{t_1 t_2}, \ldots, m_{t_u l_i}$ are nonzero.

Theorem 4: If and only if an n-order irreducible matrix M

(1) is hyperdominant, and

(2) contains no principal submatrix of a $K_{3,3}$ or K_5 matrix, then

(1) M is an ocim of a nonseparable nonredundant planar graph G_1, of nullity n, 2-isomorphic to a graph G possessing a cotree, G_c, which is a window of G with the orientation of all chords in the same direction as the window.

(2) $M = BRB'$ where B is an f-circuit matrix of G defined by the chords of G_c.

The proof of Theorem 4 follows directly from the definitions of $K_{3,3}$ and K_5 matrices.

Theorem 5 gives a synthesis technique for a matrix with some negative entries.

Theorem 5: Suppose $M = [m_{ij}]_n$ and $m_{pq}\, m_{pj}\, m_{qj} < 0$ for all j, $j \neq p, q$.

Let $m_{pq} < 0$ and define

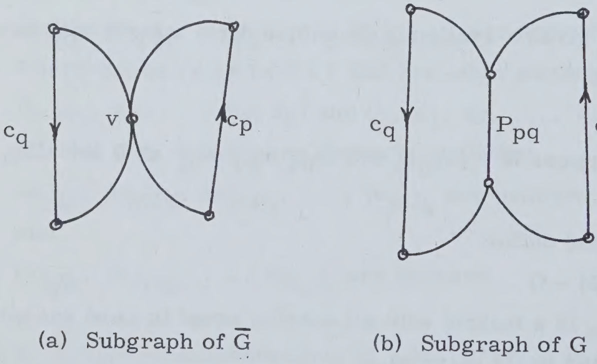
$$\overline{M} = M - \Omega$$

where $\Omega = [\omega_{ij}]_n$ is a matrix with all entries equal to zero except $\omega_{pp} = \omega_{qq} = |m_{pq}|$ and $\omega_{pq} = \omega_{qp} = m_{pq}$.

If and only if M is an ocim of a graph G of nullity n then $\overline{M}$ is an ocim of a graph $\overline{G}$ of nullity n such that the f-circuits of $\overline{G}$ defined by chords c_p and c_q have one and only one vertex in common and if c_p is similarly (or oppositely) oriented to c_j in some cutset of G for every j, so is c_q.

Proof: Sufficiency: Since M is an ocim of a graph G of nullity n and $m_{pq}\, m_{pj}\, m_{qj} < 0$ for all j, $j \neq p$, q, P_{pq} has no common element with any path P_{ij}, $i \neq p$ or q, hence $\overline{M}$ is an ocim of a graph $\overline{G}$ obtained from G by shorting P_{pq}. The nullity of $\overline{G}$ is the same as that of G. Since the f-circuits defined by c_p and c_q have common elements in G, namely P_{pq}, as a consequence of shorting P_{pq}, these two circuits have one and only one common vertex in $\overline{G}$. Since m_{pj} and m_{qj} have the same sign for all j, if c_p is similarly (or oppositely) oriented to c_j in some cutset of G, so is c_q.

Necessity: Since for every j, whenever c_p is similarly (or oppositely) oriented to c_j in some cutset of G, so is c_q; and the f-circuits of $\overline{G}$ defined by chords c_p and c_q have one and only one vertex in common, it is always possible to construct G from $\overline{G}$ such that P_{pq} has no common element with any path P_{ij} in G. A method for obtaining G from $\overline{G}$ is shown in Figure 1. Figure 1(a) is a subgraph of $\overline{G}$ consisting of f-circuits defined by c_p and c_q with a common vertex v. Figure 1(b) shows the process of splitting the vertex v and inserting path P_{pq} to obtain G.



(a) Subgraph of $\overline{G}$ (b) Subgraph of G

Figure 1. An Illustration for the Proof of Theorem 5

The necessary and sufficient conditions for synthesis of a matrix as an ocim of a graph of which all chords are incident to a common vertex (a generalized star cotree) is given next. This class of matrices is a special class of uniformly tapered matrices.

Definition 3: W-matrix: The matrix, $W = [w_{ij}]$ is a symmetric matrix obtained from a uniformly tapered matrix $M = [m_{ij}]$ by

$$w_{ij} = m_{ij} + m_{i-1, j+1} - m_{i-1, j} - m_{i, j+1} \text{ for all } i \geqslant j$$

where $m_{ij} = 0$ if either index i or j is zero or exceeds the order of M.

Definition 4: S-uniformly tapered matrix: A uniformly tapered matrix which possesses properties 1 and 2 below is called a S-uniformly tapered matrix.

Property 1: If $w_{1j_k} \neq 0$ where $1 < j_1 < j_2 < \cdots < j_k < \cdots < j_r \leqslant n$ where n is the order of W, then $w_{pq} = 0$ for $1 < p \leqslant j_k$, $j_k < q \leqslant n$, $k = 1, 2, \ldots, r$.

It is obvious that if W satisfies property 1, then $\overline{W}$, the matrix obtained from W by omitting column 1 and row 1 of W, is block diagonal. There are r blocks if $j_r = n$ and r + 1 blocks if $j_r < n$. Denote each matrix block by W_i.

Property 2: Either W_i is a 1 x 1 matrix or W_i possesses Property 1.

Theorem 6: If and only if an nth order irreducible matrix, M, is S-uniformly tapered, then M is an ocim of a graph G_1, 2-isomorphic to a nonseparable planar graph G which possesses a generalized star cotree with all chords oriented towards (or away from) the common vertex.

The proof of Theorem 6 is given elsewhere [2], hence it is omitted.

A special case of Theorem 6 is described in Corollary 2.

Corollary 2: Suppose $M = [m_{ij}]_n$ is an ocim of a graph G of nullity n with chords c_i, $i = 1, 2, \ldots, n$ defining f-circuits. If chords c_p, c_q, c_r are incident to a common vertex and oriented towards (or away from) the common vertex in G, then the entries in M_{11}, the principal submatrix of M consisting of rows p, q and r and the corresponding columns, are non-negative and satisfy either of the following conditions:

$$\left[\begin{array}{ccccc} x & \geq & x & \geq & x \\ & & \rotatebox{90}{$\geq$} & & \Vert \\ x & & x & \geq & x \\ & & & & \rotatebox{90}{$\geq$} \\ x & & x & & x \end{array}\right] \quad \text{or} \quad \left[\begin{array}{ccccc} x & \geq & x & = & x \\ & & \rotatebox{90}{$\geq$} & & \rotatebox{90}{$\geq$} \\ x & & x & \geq & x \\ & & & & \rotatebox{90}{$\geq$} \\ x & & x & & x \end{array}\right]$$

In the above pictorial representation, x is an entry.

CONCLUSIONS

A number of new topological properties of open circuit impedance matrices are presented. Conditions on the signs and the magnitude of the entries in principal submatrices of n-order ocim for a graph of nullity n are determined. The necessary and sufficient conditions for the synthesis of matrices as ocim of two classes of planar networks are given. Also it is shown that some sign and magnitude constraints on the entries in the principal submatrices require certain topological properties in the corresponding graph.

REFERENCES

1. D. P. Brown, "Topological properties of resistance matrix," SIAM J. Applied Math, Vol. 16, No. 2, pp. 387-394, March 1968.
2. A. Y. Lee, "New necessary and sufficient conditions for impedance matrix synthesis," Ph. D. Dissertation, University of Wisconsin, 1969.
3. D. P. Brown, "Synthesis of a class of planar n-ports," Trans. on Circuit Theory, Vol. CT-12, pp. 612-613, Dec. 1965.

STRUCTURE OF DECAYING SOLUTIONS OF SINGULAR PERTURBATION PROBLEMS

FRANK HOPPENSTEADT *
Courant Institute of Mathematical Sciences, New York University,
New York, New York 10012

Abstract. Systems of the form $dx/dt = f(t,x,y,\varepsilon)$, $\varepsilon\, dy/dt = g(t,x,y,\varepsilon)$ are considered. This paper studies conditions under which the solution of the initial value problem associated with this system can be written as $\{x^*(t,\varepsilon), y^*(t, \varepsilon)\} + \{X(\tau,\varepsilon), Y(\tau,\varepsilon)\}$ where (x^*,y^*) is a solution of the system which is smooth at $\varepsilon = 0$ and (X,Y) is an exponentially decaying function of $\tau = t/\varepsilon$ as $\tau \to \infty$.

Consider the system of scalar equations

$$\begin{aligned} dx/dt &= f(x,y,\varepsilon), \qquad x(0) = \xi(\varepsilon), \\ \varepsilon\, dy/dt &= g(x,y,\varepsilon), \qquad y(0) = \eta(\varepsilon), \end{aligned} \tag{P_ε}$$

where ε is a real parameter near zero. We assume the reduced problem

$$\begin{aligned} dx/dt &= f(x,y,0), \qquad x(0) = \xi(0), \\ 0 &= g(x,y,0) \end{aligned} \tag{P_0}$$

has a solution $x = x_0(t)$, $y = y_0(t)$, on some interval $0 \leq t \leq T$. We also assume f,g have continuous derivatives for (x,y) near (x_0,y_0) and $|\varepsilon| \leq \varepsilon_0$.

Our main assumption is

$$(\partial g/\partial y)\{x_0(t),\ y_0(t),\ 0\} < 0 \tag{J}$$

for $0 \leq t \leq T$.

With these conditions, we can show for any choice of $\xi^*(\varepsilon)$ for which $\xi^*(0) = x_0(0)$, there is a solution $x = x^*(t,\varepsilon,\xi^*)$, $y = y^*(t,\varepsilon,\xi^*)$, of

$$\begin{aligned} dx/dt &= f(x,y,\varepsilon), \qquad x(0) = \xi^*(\varepsilon), \\ \varepsilon\, dy/dt &= g(x,y,\varepsilon), \end{aligned}$$

which is a smooth function of ε at $\varepsilon = 0$ and satisfies

$$x^*(t,0,\xi^*) = x_0(t), \qquad y^*(t,0,\xi^*) = y_0(t)$$

for $0 \leq t \leq T$. Here the notation $x^*(t,\varepsilon,\xi^*)$ reflects the dependence of the solution on the choice of the function ξ^*. It can be shown that x^*,y^* are smooth functions of ξ^*.

With this result, we let $X = x - x^*$, $Y = y - y^*$ and $\tau = t/\varepsilon$ in (P_ε). The result is

---

[*] This research was sponsored in part by the National Science Foundation, GP-18682.

$$dX/d\tau = \varepsilon[f(X + x^*, Y + y^*, \varepsilon) - f(x^*,y^*,\varepsilon)],\ X(0) = \xi(\varepsilon) - \xi^*(\varepsilon),$$

$$(S_\varepsilon)$$

$$dY/d\tau = g(X + x, Y + y^*,\varepsilon) - g(x^*,y^*,\varepsilon),\ Y(0) = \eta(\varepsilon) - y^*(0,\varepsilon,\xi^*)$$

Instead of this initial value problem, we consider a special initial value problem which in integrated form is

$$X(\tau) = -\int_\tau^{T/\varepsilon} [f(X + x^*, Y + y^*, \varepsilon) - f(x^*, y^*, \varepsilon)]d\sigma$$

$$Y(\tau) = e^{\int_0^\tau g_y}[\eta(\varepsilon) - y^*(0,\varepsilon,\xi^*)] + \int_0^\tau \exp(\int_\sigma^\tau g_y^o)\{g_x^o X + R(X,Y, \varepsilon, x^*, y^*)\}d\sigma$$

where $g_x^o = (\partial g/\partial x)(t,x_o(t),y_o(t),0)$, etc. Iteration arguments show this problem has a unique solution on $0 \leq \tau \leq T/\varepsilon$ provided $\varepsilon > 0$ and $|\eta(\varepsilon) - y^*(0,\varepsilon,\xi^*)|$ are sufficiently small. This result holds in general for $\Delta = |\eta(\varepsilon) - y^*(0,\varepsilon,\xi^*)|$ sufficiently small. However, it is not necessary that Δ be small for this result to hold. In fact, the set of $\eta(\varepsilon)$ for which this system of integral equations has a solution with the properties listed is roughly determined by the domain of attraction of the rest point $Y = 0$ of $dY/d\tau = g(x_o(0), Y + y_o(0), 0)$. Finally, there are positive constants K,δ independent of ε and the initial data such that

$$|X(\tau,\varepsilon)| + |Y(\tau,\varepsilon)| \leq K|\eta(\varepsilon) - y^*(0,\varepsilon,\xi^*)|e^{-\delta\tau}$$

for $0 \leq \tau \leq T/\varepsilon$. This estimate can be used to show $\lim_{\varepsilon\to 0} (X(\tau,\varepsilon), Y(\tau,\varepsilon))$ exists uniformly for $0 \leq \tau \leq T/\varepsilon$.

Thus, (S_ε) has a unique solution on $0 \leq \tau \leq T/\varepsilon$ provided ξ, $\tilde{\xi} = \xi^*(\varepsilon) - x_o(0)$ and $\eta(\varepsilon)$ satisfy

$$\Delta(\tilde{\xi},\varepsilon) = 0$$

where

$$\Delta(\tilde{\xi},\varepsilon) = \xi(\varepsilon) - x_o(0) - \tilde{\xi}(\varepsilon) + \varepsilon\int_0^{T/\varepsilon} [f(X+x^*,Y+y^*,\varepsilon)-f(x^*,y^*,\varepsilon)]d\sigma.$$

Here

$$X = X(\sigma,\varepsilon,\xi^*), \qquad x^* = x^*(\varepsilon\sigma,\varepsilon,\xi^*), \quad \text{etc.}$$

Straightforward arguments show

$$\Delta(0,0) = 0 \quad \text{and} \quad (\partial\Delta/\partial\tilde{\xi})(0,0) = -I.$$

It follows from an implicit function theorem that $\Delta(\tilde{\xi},\varepsilon) = 0$ defines a unique $\tilde{\xi}(\varepsilon)$ for each small $\varepsilon > 0$.

The result of these remarks is that for $|\eta(\varepsilon) - y_o(0)|$ sufficiently small, (P_ε) has a unique solution on $0 \leq t \leq T$ for each small ε. Moreover, the solution can be written as

$$X(t,\varepsilon) = x^*(t,\varepsilon) + X(\tau,\varepsilon), \qquad y(t,\varepsilon) = y^*(t,\varepsilon) + Y(\tau,\varepsilon)$$

where x^*,y^* are smooth functions of ε at $\varepsilon = 0$ for $0 \leq t \leq T$ and X,Y are smooth functions of ε at $\varepsilon = 0$ for $0 \leq \tau \leq T/\varepsilon$. Finally, there are positive constants K,σ such that

$$|X(\tau,\varepsilon)| + |Y(\tau,\varepsilon)| \leq Ke^{-\sigma\tau} \qquad |\eta(\varepsilon) - y^*(0,\varepsilon)|$$

for $0 \leq \tau \leq T/\varepsilon$, $0 < \varepsilon \leq \varepsilon_o$. We refer to x^*,y^* as the outer solution of the problem and X,Y as the boundary layer solution. The term "boundary layer solution" is used instead of "inner solution" to avoid confusion when boundary value problems are considered.

Generalizations: The method described above has been extended to treat the following problems.

1. Systems: Consider the initial value problem

$$\begin{aligned} dx/dt &= f(t,x,y,\varepsilon), \qquad & x(0) &= \xi(\varepsilon), \\ \varepsilon\, dy/dt &= g(t,x,y,\varepsilon), \qquad & y(0) &= \eta(\varepsilon), \end{aligned} \tag{P^*_ε}$$

where $x,f \in E^m$, $y,g \in E^n$. Instead of (J) we now assume the matrix $(\partial g/\partial y)(t,x_o(t),y_o(t),0)$ can be transformed into its Jordan canonical form in a smooth way, and no eigenvalues vanish on $0 \leq t \leq T$.

Also, we assume

$$k_1,\ 1 \leq k_1 \leq n, \qquad \text{eigenvalues of } (\partial g/\partial y)(t,x_o(t),y_o(t),0)$$

have negative real parts, say Re $\lambda(t) \leq -\mu < 0$, while the others have Re $\lambda(t) > -\mu$. (Note the other eigenvalues may have real parts positive, negative or zero.)

With such additional conditions, we can show there is a k_1-dimensional manifold $S(\varepsilon) C E^n$ such that $\eta(\varepsilon) \in S(\varepsilon)$ implies (P^*_ε) has a unique solution on $0 \leq t \leq T$. Moreover, there is an outer solution $x^*(t,\varepsilon)$, $y^*(t,\varepsilon)$, and a boundary layer $X(\tau,\varepsilon)$, $Y(\tau,\varepsilon)$, such that

$$x(t,\varepsilon) = x^*(t,\varepsilon) + X(\tau,\varepsilon), \qquad y(t,\varepsilon) = y^*(t,\varepsilon) + Y(\tau,\varepsilon)$$

for $0 \leq t \leq T$, $0 < \varepsilon \leq \varepsilon_o$. Finally, there are constants K,δ such that

$$|X(\tau,\varepsilon)| + |Y(\tau,\varepsilon)| \leq K|\eta(\varepsilon) - y^*(0,\varepsilon)|e^{-\delta\tau}, \text{ for } 0 \leq t \leq T,$$

$0 < \varepsilon \leq \varepsilon_o$.

2. Expansions: Depending on the smoothness of the data, we can expand the functions x^*, y^*, X, Y by Taylor's theorem to give for any $N = 1,2,\ldots,$

$$x(t,\varepsilon) = \sum_{r=0}^{N} [x_r(t) + X_r(\tau)]\varepsilon^r + O(\varepsilon^{N+1}),$$

$$y(t,\varepsilon) = \sum_{r=0}^{N} [y_r(t) + Y_r(\tau)]\varepsilon^r + O(\varepsilon^{N+1}),$$

where $O(\varepsilon^{N+1})$ holds uniformly for $0 \leq t \leq T$.

On the other hand, using this form of expansion as an Ansatz, we can develop a straightforward perturbation scheme for finding the coefficients without first finding x^*, y^*, X, Y.

First, the expansion

$$\sum x_r(t)\varepsilon^r, \qquad \sum y_r(t)\varepsilon^r$$

should give a solution of (P^*_ε). This leads to a recursive system of differential equations for these coefficients. These cannot be determined at this point because the appropriate initial conditions $(\xi^*(\varepsilon))$ have yet to be determined. Thus, the full expansion is substituted into (P^*_ε) and coefficients of like powers of ε equated. The result is a sequence of equations for $X_r(\tau)$, $Y_r(\tau)$. These can then be solved subject to the conditions

$$X_r(0) + x_r(0) = \xi_r, \qquad Y_r(0) + y_r(0) = \eta_r,$$

$$\lim_{\tau \to \infty} X_r(\tau) = 0, \qquad \lim_{\tau \to \infty} Y_r(\tau) = 0$$

for $r = 0,1,2, \ldots$. Here ξ_r, η_r are the coefficients of ε^r in the Taylor expansions of $\xi(\varepsilon), \eta(\varepsilon)$.

3. Systems with several parameters: We consider

$$\begin{aligned} dx/dt &= f(t,x,y_1, \ldots, y_M,\varepsilon), & x(0) &= \xi(\varepsilon), \\ \varepsilon_1\, dy_1/dt &= g_1(t,x,y_1, \ldots, y_M,\varepsilon), & y_1(0) &= \eta_1(\varepsilon), \\ \varepsilon_1\cdots\varepsilon_M\, dy_M/dt &= g_M(t,x,y_1, \ldots, y_M,\varepsilon), & y_M(0) &= \eta_M(\varepsilon), \end{aligned}$$

where $x, f \in E^m$, $y_j, g_j \in E^{n_j}$ and the parameters $\varepsilon_1, \ldots, \varepsilon_M \to 0$. Here we have written $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_M)$. This convention is used repeatedly here.

We now assume for each $j = 1, \ldots, M$, the Jacobian matrix

$$(\partial g_j/\partial y_j)(t,x_o(t),y_o(t),0)$$

has k_j, $1 \le k_j \le n_j$, eigenvalues with Re $\lambda(t) \le -\mu < 0$ and the rest with Re $\lambda(t) > -\mu$ for $0 \le t \le T$.

New independent variables are introduced successively to the problem. Let $\tau_q = t/\varepsilon_1 \ldots \varepsilon_q$, $q = 1, \ldots, M$. The resulting equations can be solved provided the initial data is suitably restricted.

The main result is the existence of manifolds $S_1(\varepsilon), \ldots, S_M(\varepsilon)$ where $S_i \subset E^{n_i}$ has dimension k_i, such that the problem has a unique solution for $0 \le t \le T$ and small $|\varepsilon|$ with $\varepsilon_i > 0$ provided

$$\eta_i(\varepsilon) \in S_i(\varepsilon), \qquad i = 1, \ldots, M.$$

Moreover, there is a particular outer solution $x^*(t,\varepsilon)$, $y_j^*(t,\varepsilon)$, and intermediate solutions $X^r(\tau_r,\varepsilon)$, $Y_j^r(\tau_r,\varepsilon)$ such that

$$x(t,\varepsilon) = x^*(t,\varepsilon) + \sum_{r=0}^{M} X^r(\tau_r,\varepsilon)$$

$$y(t,\varepsilon) = y^*(t,\varepsilon) + \sum_{r=0}^{M} Y^r(\tau_r,\varepsilon)$$

for $0 \le t \le T$. As in Section 2, an expansion for these functions can be found if the data is smooth.

4. $T = \infty$: In this case, some stability condition is needed for the unperturbed (x) variables. We assume

the linear system

$$\text{(L)} \quad d\phi/dt = (f_x^o(t) - f_y^o(t) g_y^{o-1}(t) g_x^o(t))\phi$$

possesses an exponential dichotomy. Moreover, we assume f,g have bounded derivatives in a neighborhood of $(t, x_o(t), y_o(t), 0)$, $0 \le t < \infty$.

Then if $\tilde{m}$, $1 \le \tilde{m} \le m$, solutions of (L) decay as $t \to \infty$, there is an $\tilde{m} + k$-manifold $S(\varepsilon) \in E^{m+n}$ such that (P_ε^*) has a unique solution for $0 \le t < \infty$ provided $(\xi(\varepsilon), \eta(\varepsilon)) \in S(\varepsilon)$. As before, this solution can be written as the sum of an outer solution and a boundary layer solution. Also, expansions can be found.

5. Boundary value problems. Consider the boundary value problem

$$\begin{aligned} dx/dt &= f(t,x,y,\varepsilon), \\ \varepsilon\, dy/dt &= g(t,x,y,\varepsilon), \end{aligned} \qquad (E_\varepsilon)$$

$$x(0) = A, \qquad x(1) = B \qquad \text{(BC)}$$

where x,y are real scalars and $0 \le t \le 1$. We assume

the reduced problem

$$dx/dt = f(t,x,y,\varepsilon), \qquad x(1) = B,$$

$$0 \quad = g(t,x,y,0)$$

has a smooth solution $x = x_o(t)$, $y = y_o(t)$ for $0 \leq t \leq T$. Also, f,g are smooth functions for $0 \leq t \leq 1$, (x,y) near $(x_o(t), y_o(t))$ and $|\varepsilon| \leq \varepsilon_o$.

We also assume

$$(\partial g/\partial y)(t,x_o(t),y_o(t),0) < 0 \qquad \text{for } 0 \leq t \leq 1.$$

With no loss of generality we may assume $(\partial g/\partial x)(t,x_o(t),y_o(t),0) \equiv 0$.

From earlier work, there is an outer solution; i.e., a solution $x = x^*(t,\varepsilon)$, $y = y^*(t,\varepsilon)$, of (E_ε) which satisfies

$$x^*(1,\varepsilon) = B \;\text{ and }\; |x^*(t,\varepsilon) - x_o(t)| + |y^*(t,\varepsilon) - y_o(t)| = 0(\varepsilon)$$

where $0(\varepsilon)$ holds uniformly for $0 \leq t \leq 1$.

We now investigate the existence of a solution of (E_ε), (BC). First, let $X = x - x^*$, $Y = y - y^*$ and $\tau = t/\varepsilon$ in (E_ε). The result is the "stretched" system

$$\begin{aligned} dX/d\tau &= \varepsilon F(\varepsilon\tau, X, Y, \varepsilon), \\ dY/d\tau &= G(\varepsilon\tau, X, Y, \varepsilon), \end{aligned} \qquad (S_\varepsilon)$$

where $F = f(\varepsilon\tau, X + x^*(\varepsilon\tau,\varepsilon), Y + y^*(\varepsilon\tau,\varepsilon),\varepsilon) - f(\varepsilon\tau,x^*(\varepsilon\tau,\varepsilon), y^*(\varepsilon\tau,\varepsilon),\varepsilon)$, etc. Thus, the boundary value problem has a solution if and only if (S_ε) has a solution whose x-components satisfies

$$X(0,\varepsilon) = A - x^*(0,\varepsilon) \text{ and } X(1/\varepsilon,\varepsilon) = 0.$$

The second of these conditions leads to consideration of the system of integral equations

$$\begin{aligned} X(\tau,\varepsilon) &= - \varepsilon \int_\tau^{1/\varepsilon} F(\varepsilon\sigma, X, Y, \varepsilon)d\sigma \\ Y(\tau,\varepsilon) &= \exp\{g_y^o\tau\}\eta + \int_0^\tau \exp\{g_y^o(\tau - \sigma)\} \mathcal{G}(\varepsilon\sigma, X, Y, \varepsilon)d\sigma \end{aligned} \qquad (I_\varepsilon)$$

where $g_y^o = (\partial g/\partial y)(0,x^*(0,\varepsilon)y^*(0,\varepsilon),0)$ and $\mathcal{G} = G - g_y^o Y$. Since $\mathcal{G}$ satisfies $\mathcal{G} = 0(|X| + |Y| + \varepsilon)(|X| + |Y|)$, a straightforward iteration shows that (I_ε) has a unique solution. Moreover, the solution, $X(\tau,\varepsilon;\eta)$, $Y(\tau,\varepsilon;\eta)$ satisfies

$$|X(\tau,\varepsilon)| + |Y(\tau,\varepsilon)| \leq Ke^{-\delta\tau}|\eta|$$

for some positive constants K,δ independent of ε(but not of η as $|\eta| \to \infty$).

It follows for each choice of $Y(0) = \eta$, (I_ε) has a unique solution which satisfies $X(1/\varepsilon,\varepsilon) = 0$. It follows that (S_ε) has a solution satisfying $X(1/\varepsilon,\varepsilon)=0$ if and only if $(X(0), Y(0))$ lies initially on the one dimensional manifold

$$\mathcal{S}^*(\varepsilon) = \{(X(0,\varepsilon;\eta),\eta)\} .$$

Therefore, the boundary value problem has a solution if and only if the set

$$\mathcal{S}(\varepsilon) = \{(X(0,\varepsilon;\eta) + x^*(0,\varepsilon),\ \eta + y^*(0,\varepsilon))\}$$

intersects the set $\{(A,y): y \in E^1\}$.

We note that $X(0,\varepsilon;\eta)$ is defined in general for small $|\eta|$ and ε. However, in many cases the domain of definition is a large set although actually finding it may be very difficult.

The result of these remarks is that the boundary value problem has a solution if and only if $X(0,\varepsilon;\eta)=A+x^*(0,\varepsilon)$ has at least one solution for η.

The following two examples illustrate that this system of equations may or may not have solutions.

Example 1: $\varepsilon d^2z/dt^2 + dz/dt + (dz/dt)^3 = 0$, $z(0) = 1$, $z(1) = 0$. We set $x = z$, $dx/dt = y$ and get

$$dx/dt = y, \quad dy/dt = -y-y^3, \; x(0) = 1, \; x(1) = 0.$$

Here the outer solution is chosen as $x^* = 0$, $y^* = 0$. Then with $X = x$, $Y = y$, $\tau = t/\varepsilon$, this system becomes

$$dX/d\tau = \varepsilon Y, \; dY/d\tau = -Y-Y^3 .$$

It follows that $\mathcal{S}(\varepsilon) = \{(-\varepsilon \sin^{-1}(\eta/\sqrt{1+\eta^2}) + 0(e^{-1/\varepsilon}),\eta)\}$ which lies in a strip about $x = 0$ of radius $0(\varepsilon)$. Hence, for small ε, the problem can have no solution.

Example 2: $\varepsilon d^2z/dt^2 + f(z,\varepsilon)(dz/dt) = 0$, $z(0) = 1$, $z(1) = 0$ where we assume the smooth function f satisfies $M \geq f(z,\varepsilon) \geq m > 0$ for some positive constants M,m and all $|z| < \infty$, $|\varepsilon| \leq \varepsilon_0$. Setting $x = z$, $dx/dt = y$, we obtain the problem $dx/dt = y$, $\varepsilon dy/dt = -f(x,\varepsilon)y$, $x(0) = 1$, $x(1) = 0$. In this case, the outer solution is $x^* = 0$, $y^* = 0$. Then with $X = x$, $Y = y$ and $\tau = t/\varepsilon$, we have

$$X(0,\varepsilon,\eta) = -\varepsilon \int_0^{1/\varepsilon} Y(\sigma,\varepsilon,\eta)d\sigma = -\varepsilon\eta \int_0^{1/\varepsilon} \exp[-\int_0^{\sigma} f(X,\varepsilon)d\sigma]d\sigma.$$

We wish to solve $X(0,\varepsilon,\eta) = 1$. This can be done since

$$(\varepsilon\eta/M)(1 - e^{-M/\varepsilon}) \leq -X(0,\varepsilon,\eta) \leq (\varepsilon\eta/m)(1 - e^{-m/\varepsilon}).$$

Moreover, $m \leq -\varepsilon\eta \leq M$. Hence, $\eta = 0(1/\varepsilon)$ and $\eta \to \infty$ as $\varepsilon \to 0$.

SINGULAR PERTURBATION THEORY OF LINEAR STATE REGULATORS

PETAR V. KOKOTOVIĆ and RICHARD A. YACKEL
Coordinated Science Laboratory and
Department of Electrical Engineering
University of Illinois
Urbana, Illinois

ABSTRACT

This paper analyzes the behavior of the solution of the Riccati equation for the linear regulator problem with a parameter whose perturbation changes the order of the system. Sufficient conditions are given under which the solution of the original problem tends to the solution of a low-order problem. This result can be used for the decomposition of a high-order problem into two low-order problems.

INTRODUCTION

In this paper we analyze the dependence of the solution of the Riccati equation for the linear state regulator problem on a parameter whose small perturbation can change the order of the system. The system considered is

$$\frac{dx}{dt} = \widetilde{A}_1 x + \widetilde{A}_2 z + \widetilde{B}_1 u, \qquad x(t_o) = x^o \tag{1a}$$

$$\lambda \frac{dz}{dt} = \widetilde{A}_3 x + \widetilde{A}_4 z + \widetilde{B}_2 u, \qquad z(t_o) = z^o \tag{1b}$$

where λ is a small positive scalar, x and z are n- and m-dimensional state vectors, respectively, u is an r-dimensional control vector, and

(i) $\widetilde{A}_1$, $\widetilde{A}_2$, $\widetilde{A}_3$, $\widetilde{A}_4$, $\widetilde{B}_1$ and $\widetilde{B}_2$ are of class C^0 in t and λ for $t \in [t_o, t_f]$, $\lambda \in [0, \lambda^o]$.

The performance index is

$$J = \frac{1}{2} y'(t_f)\widetilde{\Pi} y(t_f) + \frac{1}{2}\int_{t_o}^{t_f} (y'\widetilde{Q}y + u'\widetilde{R}u)dt \tag{2}$$

where prime denotes transpose,

This work was supported in part by the U. S. Air Force under Grant AFOSR-68-1579C, in part by the Joint Services Electronics Program (U. S. Army, U. S. Navy, and U. S. Air Force) under Contract DAAB-07-67-C-0199, and in part by the National Science Foundation under Grant GK-3893, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign.

$$y = \begin{bmatrix} x \\ z \end{bmatrix}, \quad \tilde{Q} = \begin{bmatrix} \tilde{Q}_1 & \tilde{Q}_2 \\ \tilde{Q}_2' & \tilde{Q}_3 \end{bmatrix}, \quad \tilde{\Pi} = \begin{bmatrix} \Pi_1 & \lambda\Pi_2 \\ \lambda\Pi_2' & \lambda\Pi_3 \end{bmatrix}, \tag{3}$$

and

(ii) $\tilde{Q}$ and $\tilde{R}$ are of class C^2 in t for $t \in [t_o, t_f]$ and of class C^0 in λ for $\lambda \in [0, \lambda^o]$;

(iii) $\tilde{R}$ is symmetric, positive definite, and $\tilde{Q}$ and $\tilde{\Pi}$ are symmetric, positive semidefinite for $t \in [t_o, t_f]$, $\lambda \in [0, \lambda^o]$;

(iv) Π_1, Π_2, and Π_3 are independent of λ.

The dependence of this problem on λ is not "regular" at $\lambda = 0$ since the matrices

$$\tilde{A} = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 \\ \tilde{A}_3/\lambda & \tilde{A}_4/\lambda \end{bmatrix} \qquad \tilde{B} = \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2/\lambda \end{bmatrix} \tag{4}$$

of the state equation

$$\frac{dy}{dt} = \tilde{A}y + \tilde{B}u, \qquad y(t_o) = y^o, \tag{5}$$

are not bounded at $\lambda = 0$. Hence a change of λ from $\lambda > 0$ to $\lambda = 0$ is often called a "singular perturbation." For $\lambda > 0$ the optimum control is [1]

$$u = -\tilde{R}^{-1}\tilde{B}'\tilde{K}y, \tag{6}$$

where $\tilde{K}$ is the solution of

$$\frac{d\tilde{K}}{dt} = -\tilde{K}\tilde{A} - \tilde{A}'\tilde{K} + \tilde{K}\tilde{B}\tilde{R}^{-1}\tilde{B}'\tilde{K} - Q, \qquad \tilde{K}(t_f) = \tilde{\Pi}, \tag{7}$$

and the optimum regulator system is

$$\frac{dy}{dt} = (\tilde{A} - \tilde{B}\tilde{R}^{-1}\tilde{B}'\tilde{K})y. \tag{8}$$

In Theorems 1 and 2 we establish sufficient conditions under which $\tilde{K}$ tends to a limit K as the positive parameter λ tends to zero. Except for an interval about $t = t_f$ the convergence of $\tilde{K}$ to K is uniform in t. The conditions in this paper are much less restrictive than those which appear in [2], and the results of [2] are a special case of the general theory developed here. In a second paper [3] Theorems 1 and 2 and an asymptotic expansion of K form the basis for a design procedure for high order systems.

This paper makes use of results of the singular perturbation theory of Levin and Levinson [4] and Hoppensteadt [5], and in particular of the Theorems L and H given in the Appendix.

MAIN THEOREM

Partitioning $\widetilde{K}$ in nxn, nxm, and mxm arrays $\widetilde{K}_1$, $\lambda\widetilde{K}_2$ and $\lambda\widetilde{K}_3$ respectively,

$$\widetilde{K} = \begin{bmatrix} \widetilde{K}_1 & \lambda\widetilde{K}_2 \\ \lambda\widetilde{K}_2' & \lambda\widetilde{K}_3 \end{bmatrix}$$

and denoting $\widetilde{S}_1 = \widetilde{B}_1\widetilde{R}^{-1}\widetilde{B}_1'$, $\widetilde{S}_2 = \widetilde{B}_2\widetilde{R}^{-1}\widetilde{B}_2'$ and $\widetilde{S} = \widetilde{B}_1\widetilde{R}^{-1}\widetilde{B}_2'$, we rewrite (7) as

$$\frac{d\widetilde{K}_1}{dt} = -\widetilde{K}_1\widetilde{A}_1 - \widetilde{A}_1'\widetilde{K}_1 - \widetilde{K}_2\widetilde{A}_3 - \widetilde{A}_3'\widetilde{K}_2' + \widetilde{K}_1\widetilde{S}_1\widetilde{K}_1$$

$$+ \widetilde{K}_1\widetilde{S}\,\widetilde{K}_2' + \widetilde{K}_2\widetilde{S}'\widetilde{K}_1 + \widetilde{K}_2\widetilde{S}_2\widetilde{K}_2' - \widetilde{Q}_1, \tag{10a}$$

$$\lambda\frac{d\widetilde{K}_2}{dt} = -\widetilde{K}_1\widetilde{A}_2 - \widetilde{K}_2\widetilde{A}_4 - \lambda\widetilde{A}_1'\widetilde{K}_2 - \widetilde{A}_3'\widetilde{K}_3 + \lambda\widetilde{K}_1\widetilde{S}_1\widetilde{K}_2$$

$$+ \widetilde{K}_1\widetilde{S}\,\widetilde{K}_3 + \lambda\widetilde{K}_2\widetilde{S}'\widetilde{K}_2 + \widetilde{K}_2\widetilde{S}_2\widetilde{K}_3 - \widetilde{Q}_2, \tag{10b}$$

$$\lambda\frac{d\widetilde{K}_3}{dt} = -\lambda\widetilde{K}_2'\widetilde{A}_2 - \lambda\widetilde{A}_2'\widetilde{K}_2 - \widetilde{K}_3\widetilde{A}_4 - \widetilde{A}_4'\widetilde{K}_3$$

$$+ \lambda^2\,\widetilde{K}_2'S_1\widetilde{K}_2 + \lambda\widetilde{K}_2'\widetilde{S}\,\widetilde{K}_3 + \lambda\widetilde{K}_3S'\widetilde{K}_2 + \widetilde{K}_3\widetilde{S}_2\widetilde{K}_3 - \widetilde{Q}_3, \tag{10c}$$

with the end condition

$$\widetilde{K}_1(t_f) = \Pi_1,\quad \widetilde{K}_2(t_f) = \Pi_2,\quad \widetilde{K}_3(t_f) = \Pi_3. \tag{11}$$

If λ is set to zero, (10) formally reduces to a lower order system of equations called the "reduced system,"

$$\frac{dK_1}{dt} = -K_1(A_1 - SK_2') - (A_1 - SK_2')'K_1 + K_1S_1K_1$$

$$- K_2A_3 - A_3'K_2' + K_2S_2K_2' - Q_1, \tag{12a}$$

$$0 = K_2(S_2K_3 - A_4) - K_1A_2 - A_3'K_3 + K_1SK_3 - Q_2, \tag{12b}$$

$$0 = -K_3A_4 - A_4'K_3 + K_3S_2K_3 - Q_3, \tag{12c}$$

where the absence of a tilde above a matrix denotes the evaluation of that matrix at $\lambda = 0$. Note that the dependence of (10) on λ is not regular. As $\lambda \to 0$ the derivatives $\frac{d\widetilde{K}_2}{dt}$ and $\frac{d\widetilde{K}_3}{dt}$ may tend to infinity at $t = t_f$ since, in general, Π_2 and Π_3 are not the roots of (12b,c). Hence there is an interval of time ending at t_f where the solution of (10) differs markedly from the solution of (12). One might expect that under appropriate conditions $\tilde{K}_2$ and $\tilde{K}_3$ will rapidly change from Π_2 and Π_3 to a root of (12 b,c) in an interval about $t = t_f$. This interval is often called the "boundary layer" because of an analogy with the boundary layers in fluid dynamics, [6,7].

We seek to determine whether for small λ the solution of the reduced system (12) approximates in some sense the solution of the full system (10). Our plan is to apply Theorem L given in the Appendix to Equation (10). We first introduce the "boundary layer equation"

$$\frac{dL_2(\tau)}{d\tau} = L_2(\tau)[S_2(t)L_3(\tau) - A_4(t)] + [K_1(t)S(t) - A_3'(t)]L_3(\tau) - K_1(t)A_2(t) - Q_2(t) \tag{13a}$$

$$\frac{dL_3(\tau)}{d\tau} = -L_3(\tau)A_4(t) - A_4'(t)L_3(\tau) + L_3(\tau)S_2(t)L_3(\tau) - Q_3(t) \tag{13b}$$

with the new time variable[*] τ. In Lemmas 1 and 2 we establish conditions under which a root of (12 b,c) is an asymptotically stable equilibrium of (13), for each value of the fixed parameter t on the interval $[t_o, t_f]$.

Lemma 1

Let (i) - (iv) be satisfied and assume that for each $t \in [t_o, t_f]$ the conditions (v) and (vi) hold in either Case 1, or Case 2, or Case 3 below:

Case 1

(v) rank $[C_3', A_4' C_3', \dots. (A_4')^{m-1} C_3'] = m$, where C_3 is a solution of $C_3' C_3 = Q_3$;

(vi) rank $[B_2, A_4 B_2, \dots, (A_4)^{m-1} B_2] = m$.

Case 2

(v) Q_3 is positive definite,

(vi) there exists a real matrix D for which $A_4 + B_2 D$ is a stable[**] matrix.

Case 3

(v) $Q_3 \equiv 0$,

(vi) A_4 is a stable matrix.

Consider t as a fixed parameter. Then for each $t \in [t_o, t_f]$, first, there exists a unique isolated[†] positive definite root $\mathcal{K}(t)$ of (12c);

---

[*] This variable is often referred to as the "fast time" since (13) can be viewed as being obtained from (10) by the use of the "stretching transformation" $t = \lambda\tau + t_f$ and allowing $\lambda \to 0$, see page 254 of [6].

[**] A matrix is stable if all its eigenvalues have negative real parts.

[†] Isolated roots and the domain of attraction are defined in the usual sense as in [11].

second, this root is an asymptotically stable equilibrium of (13b) as $\tau \to -\infty$; third, Π_3 belongs to the domain of attraction[†] of the asymptotically stable equilibrium $\mathcal{K}_3(t)$ of (13b); and fourth,

$$\alpha(t) = Q_4(t) - S_2(t)\,\mathcal{K}_3(t) \tag{14}$$

is a stable matrix.

The proof is omitted since Case 1 follows directly from [1,8], Case 2 from [9,10] and Case 3 from [2].

Lemma 2

If (i) - (vi) are satisfied then the root $\mathcal{K}_2(t)$ of (12b) at $t = t_f$,

$$\mathcal{K}_2(t_f) = [\Pi_1 S(t_f)\mathcal{K}_3(t_f) - \Pi_1 A_2(t_f) - A_3'(t_f)\mathcal{K}_3(t_f) - Q_2(t_f)]\alpha^{-1}(t_f), \tag{15}$$

where $K_1(t_f) = \Pi_1$, is an asymptotically stable equilibrium of (13a) as $t \to -\infty$.

Proof: Rewrite (13a) as

$$\frac{dL_2(\tau)}{d\tau} = -L_2(\tau)[\alpha(t_f) + \beta(\tau,t_f)] + \gamma(\tau,t_f) \tag{16}$$

where

$$\beta(\tau,t_f) = S_2(t_f)[\mathcal{K}_3(t_f) - L_3(\tau)], \tag{17a}$$

$$\gamma(\tau,t_f) = -\Pi_1[A_2(t_f) - S(t_f)L_3(\tau)] - A_3'(t_f)L_3(\tau) - Q_2(t_f). \tag{17b}$$

In view of known results for the stability of linear systems,[*] it follows that if: (a) $\beta(\tau,t_f)$ is continuous in τ, (b) $\beta(\tau,t_f) \to 0$ as $\tau \to -\infty$, and (c) the null solution of the honogeneous part of (16) is asymptotically stable; then $\mathcal{K}_2(t_f)$ is an asymptotically stable equilibrium of (16). Conditions (a) and (b) are satisfied by the definition (17a) of $\beta(\tau,t_f)$. To prove (c) let $\ell_i(\tau)$ be the ith row of L_2 and consider the homogeneous part of (16)

$$\frac{d\ell_i}{d\tau} = -[A_4(t_f) - S_2(t_f)\mathcal{K}_3(t_f)]\ell_i \quad i = 1,2,\dots,n. \tag{18}$$

Since by Lemma 1 in all three cases $[A_4(t_f) - S_2(t_f)\mathcal{K}_3(t_f)]$ is a stable matrix, the null solution of (18) is an asymptotically stable equilibrium as $\tau \to -\infty$.

With Lemmas 1 and 2 we satisfy condition L2 in the Appendix. To satisfy the condition L3 we establish the existence and uniqueness of the solution $K_1(t)$, $\mathcal{K}_2(t)$ and $\mathcal{K}_3(t)$ of the reduced system (12) with $K_1(t_f) = \Pi_1$. Since $\mathcal{K}_3(t)$ is uniquely obtained from (12c), we use (12b) to express $\mathcal{K}(t)$ in terms of $K_1(t)$ and $\mathcal{K}_3(t)$,

---

[*] See Theorem 9, p. 70 of [12].

$$\mathcal{K}_2 = (K_1 S\mathcal{K}_3 - K_1 A_2 - A_3'\mathcal{K}_3 - Q_2)(A_4 - S_2\mathcal{K}_3)^{-1}, \tag{19}$$

and then substitute $\mathcal{K}_2$ into (12a):

$$\frac{dK_1}{dt} = -K_1\hat{A} - \hat{A}'K_1 + K_1\hat{B}R^{-1}\hat{B}'K_1 - \hat{Q}, \quad K_1(t_f) = \Pi_1, \tag{20}$$

where

$$\hat{A} = A_1 + E_1 A_3 + SE_2' + E_1 S_2 E_2' \tag{21a}$$

$$\hat{B} = B_1 + E_1 B_2 \tag{21b}$$

$$\hat{Q} = -E_2 A_3 - A_3' E_2' - E_2 S_2 E_2' + Q_1 \tag{21c}$$

$$E_1 = (S\mathcal{K}_3 - A_2)(A_4 - S_2\mathcal{K}_3)^{-1}, \quad E_2 = (A_3'\mathcal{K}_3 + Q_2)(A_4 - S_2\mathcal{K}_3)^{-1}. \tag{21d}$$

Lemma 3

If (i) - (vi) are satisfied, then the solution K_1 of (2) exists and is unique on the interval $[t_o, t_f]$.

Proof: It follows from [1] that if: (a) R is symmetric, positive definite, (b) $\hat{Q}$ is symmetric, positive semidefinite; then the solution K_1 exists and is unique on the interval $[t_o, t_f]$. Condition (a) is satisfied by assumption (iii). To show that condition (b) is satisfied apply a matrix identity from (12c) to (21c) to obtain

$$\hat{Q} = (Q_2 + A_3'\mathcal{K}_3)(A_4'\mathcal{K}_3 + Q_3)^{-1} Q_3 (A_4'\mathcal{K}_3 + Q_3)'^{-1}(Q_2 + A_3'\mathcal{K}_3)'$$

$$- Q_2(A_4'\mathcal{K}_3 + Q_3)'^{-1}(Q_2' + \mathcal{K}_3 A_3) - (Q_2' + \mathcal{K}_3 A_3)'(A_4'\mathcal{K}_3 + Q_3)^{-1} Q_2' + Q_1. \tag{22}$$

For an arbitrary vector θ let

$$v = \theta \text{ and } w = -(A_4'\mathcal{K}_3 + Q_3)'^{-1}(Q_2 + A_3'\mathcal{K}_3)'\theta. \tag{23}$$

Since by (iii) Q is positive semidefinite

$$\theta'\hat{Q}\theta = v'Q_1 v + v'Q_2 w + w'Q_2' v + w'Q_3 w \geq 0. \tag{24}$$

Thus $\hat{Q}$ is also positive semidefinite.

Finally we show in Lemma 4 that the Jacobian of the system (10 b,c) has the property required by the condition L4 in the Appendix. We map the matrices $\widetilde{K}_1$, $\widetilde{K}_2$ and $\widetilde{K}_3$ into nn-, nm-, and mm-vectors $\underline{K}_1$, $\underline{K}_2$ and $\underline{K}_3$ respectively. We rewrite (10) in vector form

$$\frac{d\underline{K}_1}{dt} = f_1(\underline{K}_1, \underline{K}_2, t, \lambda), \tag{25}$$

$$\lambda \frac{d\underline{K}_2}{dt} = f_2(\underline{K}_1, \underline{K}_2, \underline{K}_3, t, \lambda), \tag{26a}$$

$$\lambda \frac{d\underline{K}_3}{dt} = f_3(\underline{K}_2, \underline{K}_3, t, \lambda), \tag{26b}$$

and observe that the Jacobian $\Gamma(\lambda)$ of the subsystem (23 a,b) is

$$\Gamma(\lambda) = \begin{bmatrix} \Gamma_{22}(\lambda) & \Gamma_{23}(\lambda) \\ \Gamma_{32}(\lambda) & \Gamma_{33}(\lambda) \end{bmatrix}, \tag{27}$$

where

$$\Gamma_{i,j}(\lambda) = \frac{\partial f_i}{\partial \underline{K}_j}(\underline{K}_1, \underline{K}_2, \underline{K}_3, t, \lambda), \quad i,j = 2,3.$$

Let $\Gamma(0)$ denote $\Gamma(\lambda)$ for $\lambda = 0$ evaluated along the solution of (12).

Lemma 4

If (i) - (vi) are satisfied then all the eigenvalues of $\Gamma(0)$ have positive real parts for $t\in[t_o,t_f]$.

Proof: Since $\Gamma_{23}(0) = 0$ the eigenvalues of $\Gamma(0)$ consist of the eigenvalues of $\Gamma_{22}(0)$ and those of $\Gamma_{33}(0)$. Note that $\Gamma_{22}(0)$ and $\Gamma_{33}(0)$ can be expressed in terms of Kronecker products [12],

$$\Gamma_{22}(0) = -(A_4 - S_2K_3)\times I_N, \tag{28a}$$

$$\Gamma_{33}(0) = -I_M\times(A_4 - S_2K_3)' - (A_4 - S_2K_3)'\times I_M, \tag{28b}$$

where I_N and I_M are nxn and mxm identity matrices. Therefore, the mm eigenvalues of $\Gamma_{33}(0)$ are $\mu_i + \mu_j$, $i,j = 1,\ldots,m$, and the nm eigenvalues of $\Gamma_{22}(0)$ are μ_i, $i = 1,\ldots,m$, each of multiplicity n, where $\mu_i, i=1,\ldots,m$ are the eigenvalues of $-(A_4 - S_2K_3)$, [14]. By Lemma 1 for $K_3 = \mathcal{K}_3$ all the eigenvalues μ_i have positive real parts.

With Lemmas 1, 2, 3 and 4 we satisfy all the conditions of Theorem L in the Appendix. Thus the following Main Theorem holds.

Theorem 1

If (i) - (vi) are satisfied then for sufficiently small λ the solution $\tilde{K}(t,\lambda)$ of (10) satisfying the end condition $\tilde{K}(t_f,\lambda) = \tilde{\Pi}$ exists and is unique. Furthermore,

$$\lim_{\lambda\to 0} \tilde{K}_1(t,\lambda) = K_1(t), \quad t\in[t_o,t_f], \tag{29}$$

$$\lim_{\lambda\to 0} \tilde{K}_2(t,\lambda) = \mathcal{K}_2(t), \quad t\in[t_o,t_f), \tag{30a}$$

$$\lim_{\lambda\to 0} \tilde{K}_3(t,\lambda) = \mathcal{K}_3(t), \quad t\in[t_o,t_f). \tag{30b}$$

The limit (29) is uniform in t on the interval $[t_o,t_f]$ and the limits (30) are uniform in t on any interval $[t_o,t_f]$, where $t_o \leq t_1 < t_f$.

TIME-INVARIANT PROBLEM

An important class of regulator problems is when $t_f \to \infty$ and the system (1) and $\tilde{Q}$ and $\tilde{R}$ in (2) are time-invariant. We consider this problem and assume that:

(i*) $\tilde{A}_1$, $\tilde{A}_2$, $\tilde{A}_3$, $\tilde{A}_4$, $\tilde{B}_1$ and $\tilde{B}_2$ are time-invariant and of class C^0 in λ for $\lambda \in [0,\lambda^o]$;

(ii*) $\tilde{R}$ and $\tilde{Q}$ are time-invariant and of class C^0 in λ for $\lambda \in [0,\lambda^o]$;

(iii*) $\tilde{R}$ is symmetric positive definite and $\tilde{Q}$ is symmetric positive semi-definite for $\lambda \in [0,\lambda^o]$; and

(iv*) $\Pi_1 = \Pi_2 = \Pi_3 = 0$.

Theorem 2

Let (i*) - (iv*) and (v) - (vi) be satisfied. Also let the following conditions (vii) and (viii) in either Case 1 or Case 2 be satisfied.

Case 1

(vii) rank $[\hat{C}',\hat{A}'\hat{C}',\ldots,(\hat{A}')^{n-1}\hat{C}'[= n$ where $\hat{C}$ is a solution of $\hat{C}'\hat{C} = \hat{Q}$

(viii) rank $[\hat{B},\hat{A}\hat{B},\ldots,(\hat{A})^{n-1}\hat{B}] = n$.

Case 2

(vii) $\hat{Q}$ is positive definite[*]

(viii) there exists a real matrix $\hat{D}$ for which $\hat{A} + \hat{B}\hat{D}$ is a stable matrix.

Then for sufficiently small λ the solution of the full system (10) exists for $t_o \leq t < \infty$, and the limits (26) and (27), where K_1 is the positive definite root of

$$0 = -K_1\hat{A} - \hat{A}'K_1 + K_1\hat{B}R^{-1}\hat{B}'K_1 - \hat{Q}, \tag{31}$$

hold uniformly on all closed subintervals of $t_o < t < \infty$.

Proof: If conditions (i*) - (iv*) and (v) - (vi) are satisfied then by Theorem 1 there exists a $\lambda^* > 0$ and a $t^* < t_f$ such that[†] $\|\tilde{K}(t^*,\lambda^*) - K(t^*)\| < \epsilon$ for an ϵ which satisfies the ϵ-closeness requirement of Theorem H. Furthermore, (10) has the unique solution $\tilde{K}(t,\lambda^*)$ on $[t^*,t_f]$ satisfying the end condition $\tilde{K}(t_f,\lambda^*) = 0$ and $\tilde{K}(t,\lambda) \to K(t)$ as $\lambda \to 0$ on $[t^*,t_1]$ where $t^* < t_1 < t_f$.

---

[*] Note that by (30) $\hat{Q}$ is positive definite if Q is positive definite.

[†] The norm of a matrix is taken to be the sum of the absolute values of its elements.

Hypotheses H2 - H5 of Theorem H are evidently satisfied by the form of (10) and conditions (i*) - (iii*). From the results of the linear regulator theory, conditions (i*) - (iv*) and (vii) - (viii) insure the existence and asymptotic stability of the solution K(t) of (20) on $(-\infty, t^*]$, as required by H1 and H6. The crucial hypothesis H7 is that the solution of the boundary layer equation (13) be asymptotically stable uniformly in initial conditions and the parameters t and K_1 for $t \in (-\infty, t^*]$ and K_1 positive definite. By (i*) and (ii*) the system (13) does not depend on t, nor does (13b) contain K_1. Hence by Lemma 1 the solution L_3 of (13b) satisfies H7. By Lemma 2 the solution L_2 of (13a) is asymptotically stable uniformly in K_1.

The above proof can also be based on the work of A. B. Vasil'eva [15].

CONCLUDING REMARKS

In physical systems our parameter λ represents small time constants, masses, moments of inertia etc.* There are several practical reasons which urge a designer to neglect these small parameters during the design of a regulator system. An evident reason is that the presence of these parameters can make the dimensionality of a dynamic system prohibitively high. Another, less apparent, reason is that the equations describing systems with small parameters belong to a class of "stiff" differential equations, which are difficult to solve even when the dimensionality of the system is not high.

The designer can choose one of several methods to neglect λ. The method outlined here is to neglect λ in the Riccati equation (10) and to obtain the regulator matrix from (12). Computationally an $\frac{1}{2}(n+m+1)(n+m)$-dimensional "stiff" differential system (10) is decomposed in one $\frac{1}{2}(m+1)m$-dimensional algebraic system (12c) and one $\frac{1}{2}(n+1)n$-dimensional differential system (20), with intermediate algebraic operations (19) and (21).

The results of this paper form a rigorous basis for the above method. In view of Theorems 1 and 2, K_1, $\mathcal{K}_2$ and $\mathcal{K}_3$ are an approximation of $\tilde{K}_1$, $\tilde{K}_2$ and $\tilde{K}_3$ at least for sufficiently small λ. A useful property of this approximation is that for time-invariant systems $\mathcal{K}_3$ is time-invariant even if t_f is finite. For larger values of λ an asymptotic expansion method is developed in [3].

Another method of reducing the system order is to neglect λ in the system (1) and, if A_4 has an inverse, to substitute $z = -A_4^{-1}(A_3x + B_2u)$ into (1a) and (2). If the weighting matrix for x, obtained after the substitution of z into (2), is positive semidefinite, then the Riccati formalism can be used to design a regulator for the reduced n-dimensional model.

$$\frac{dx}{dt} = (A_1 - A_2A_4^{-1}A_3)x + (B_1 - A_2A_4^{-1}B_2)u \tag{32}$$

---

*For example, if T is a small time constant and M is a small mass, then we can write $T = q_1\lambda$ and $M = q_2\lambda$, where q_1 and q_2 are appropriate coefficients.

of the system (1). Note that in this method the only state variables employed for feedback are the x-variables. In some special cases, such as Case 3 of Lemma 1, this second method coincides with the method in this paper. Further study is needed to determine whether it can be justified for a broader class of problems.

ACKNOWLEDGMENT

The authors thank Messrs. C. Hadlock, and G. Singh and Professors J. B. Cruz, Jr., B. C. Kuo, and W. R. Perkins for helpful discussions.

REFERENCES

1. R. E. Kalman, "Contributions to the Theory of Optimal Control," Bol. Socieda Mathematica, pp. 102-119, 1960.

2. P. Sannuti and P. Kokotović, "An Approximation Design of Optimal Regulators for High-Order Linear Plants," IEEE Trans. Automatic Control, AC-13, pp. 15-21, February, 1968.

3. R. Yackel and P. Kokotović, "Singular Perturbation Theory of Linear State Regulators, Part Two: Asymptotic Expansions," to be published.

4. J. Levin and N. Levinson, "Singular Perturbation of Nonlinear Systems of Differential Equations and an Associated Boundary Layer Equation," J. of Rational Mechanics and Analysis, 3, pp. 247-270, 1954.

5. F. C. Hoppensteadt, "Singular Perturbations on the Infinite Interval," Trans. AMS, 123, 2, pp. 521-535, 1966.

6. W. R. Wasow, Asymptotic Expansions for Ordinary Differential Equations, New York: Interscience, 1965.

7. M. VanDyke, Perturbation Methods in Fluid Mechanics, New York: Academic Press, 1964.

8. J. B. Moore and B. D. O. Anderson, "Optimal Linear Control Systems with Input Derivative Constraints," Proc. IEE, 114, 12, 1987-1990, December, 1967.

9. D. L. Lukes, "Optimal Regulation of Nonlinear Dynamical Systems," SIAM Journal on Control, 7, 1, pp. 75-100, February, 1969.

10. E. B. Lee and L. Markus, Foundations of Optimal Control Theory, New York: John Wiley & Sons, Inc., 1967.

11. W. Hahn, Stability of Motion, New York: Springer-Verlag, 1967.

12. W. A. Coppel, Stability and Asymptotic Behavior of Differential Equations, Boston: D. C. Heath and Co., 1965.

13. T. R. Blackburn, "Solution of the Algebraic Matrix Riccati Equation via Newton-Raphson Iteration," 1968 JACC Preprints, pp. 940-945, Ann Arbor, Michigan.

14. R. Bellman, Introduction to Matrix Analysis, New York: McGraw-Hill, 1960.

15. A. B. Vasil'eva, "Asymptotic Formulas Valid on a Half Line for Solutions of Ordinary Differential Equations with a Small Parameter in the Term Involving the Highest Derivative," Soviet Math., 3, pp. 147-150, 1962.

APPENDIX

Consider the initial value problem

$$\frac{d\xi}{dt} = f(t,\xi,\eta,\lambda), \quad \xi(t_o) = \xi_o \tag{A1}$$

$$\lambda \frac{d\eta}{dt} = g(t,\xi,\eta,\lambda), \quad \eta(t_o) = \eta_o \tag{A2}$$

where λ is a small positive parameter and ξ and η are n- and m-dimensional vectors respectively. Formally setting $\lambda = 0$ in the "full system" (A1) - (A2) gives the "reduced system"

$$\frac{d\bar{\xi}}{dt} = f(t,\bar{\xi},\bar{\eta},0), \quad \bar{\xi}(t_o) = \xi_o \tag{A3}$$

$$0 = g(t,\bar{\xi},\bar{\eta},0). \tag{A4}$$

Introducting a new time variable τ (A2) is rewritten in the form of a "boundary layer" system,

$$\frac{d\sigma}{d\tau} = g(\alpha,\beta,\sigma,0), \quad \sigma(0) = \eta_o \tag{A5}$$

where $\alpha = t_o$ and $\beta = \xi_o$ are fixed parameters. In the space of variables ξ, η, σ, t we define a region $\mathcal{R}$: $\|\xi - \bar{\xi}(t)\| < r$, $\|\eta - \bar{\eta}(t)\| < r$, $0 \leq \lambda \leq \lambda^o$, $t_o \leq t \leq t_f$ where $r > |\eta_o - \bar{\eta}(t_o)| \geq 0$.

Theorem L [4]

Let the following conditions be satisfied:

L1 $f, \frac{\partial f}{\partial \xi}, \frac{\partial f}{\partial \eta}, g, \frac{\partial y}{\partial \xi}, \frac{\partial g}{\partial \eta}$ are of class C^0 in $(\xi,\eta,,t,\lambda)\in\mathcal{R}$.

L2 The solution of $\sigma(\tau)$ of (A5) exists on $\tau\in[0,\infty)$, is unique and is asymptotically stable with respect to the isolated root $\sigma = \varphi(t_o,\xi_o)$ of $g(\alpha,\beta,\sigma,0) = 0$.

L3 The solution $\bar{\xi}(t)$, $\bar{\eta}(t)$ of (A3) - (A4) exists and is unique on $t\in[t_o,t_f]$.

L4 The real parts of the eigenvalues of the Jacobian matrix

$$\frac{\partial g}{\partial \eta}(t,\bar{\xi}(t), \bar{\eta}(t), 0) \tag{A6}$$

are negative on $[t_o,t_f]$.

Then for sufficiently small λ, the full system (A1) - (A2) has a unique solution $\xi(t,\lambda)$, $\eta(t,\lambda)$ on $t\in[t_o,t_f]$ satisfying the initial conditions $\xi(t_o,\lambda) = \xi_o$, $\eta(t_o,\lambda) = \eta_o$. Furthermore,

$$\lim_{\lambda \to 0} \xi(t,\lambda) = \bar{\xi}(t) \quad \text{on} \quad [t_o,t_f] \tag{A7}$$

$$\lim_{\lambda \to 0} \eta(t,\lambda) = \bar{\eta}(t) \quad \text{on} \quad (t_o,t_f] \tag{A8}$$

where the limit (A7) is uniform in t on $[t_o,t_f]$ and the limit (A8) is uniform in t on any interval $[t_1,t_f]$, where $t_o < t_1 < t_f$.

Theorem H [5]

Let $t_f = \infty$ in the definition of $\Re$, and denote by Ω its (n+1)-dimensional subspace in ξ and t. Let the following conditions be satisfied:

H1 The solution $\bar{\xi}(t)$, $\bar{\eta}(t)$ of the reduced system (A3) - (A4) exists on $[t_o,\infty)$.

H2 f, g, $\frac{\partial f}{\partial \xi}$, $\frac{\partial f}{\partial \eta}$, $\frac{\partial g}{\partial t}$, $\frac{\partial g}{\partial \xi}$, $\frac{\partial g}{\partial \eta}$ are of class C^1 for all $(t,\xi,\eta,\lambda)\in\Re$.

H3 There exists a bounded function, $\eta = \varphi(\xi,t)$, of class C^2 which is an isolated root of $g(t,\xi,\varphi(t,\xi),0) = 0$ for all $(\xi,t)\in\Omega$.

H4 The function f is of class C^0 $\eta(t) = \bar{\eta}(t)$, $\lambda = 0$ uniformly in $(\xi,t)\in\Omega$ and $f(t,\xi,\bar{\eta}(t),0)$ and $\frac{\partial f}{\partial \xi}(t,\xi,\bar{\eta}(t),0)$ are bounded in Ω.

H5 The function g is of class C^0 at $\lambda = 0$ uniformly in $(t,\xi,\eta)\in\Re$ and g, $\frac{\partial g}{\partial t}$, $\frac{\partial g}{\partial \xi}$, $\frac{\partial g}{\partial \eta}$ at $\lambda = 0$ are bounded on $\Re$.

H6 The solution of the reduced system is uniformly asymptotically stable.

H7 The solution of the boundary layer system is uniformly asymptotically stable uniformly in the parameters $(\alpha,\beta)\in\Omega$.

If the ϵ-closeness requirement $\|\eta_o - \varphi(t_o,\xi_o)\| < \epsilon$ is satisfied for a sufficiently small ϵ, then for a sufficiently small λ the solution of the full system exists for $t_o \leq t < \infty$. Furthermore the limits (A7) and (A8) hold uniformly on all closed subsets of $t_o < t < \infty$.

SINGULAR PERTURBATION OF A NONLINEAR BOUNDARY VALUE PROBLEM

R. E. O'MALLEY, JR.

Courant Institute of Mathematical Sciences and Department of Mathematics, University Heights, New York University, New York, New York

ABSTRACT

Asymptotic expansions for solutions of two-point boundary value problems for the system

$$\frac{dx}{dt} = f_1(x,t;\varepsilon) + f_2(x\ ,t;\varepsilon)y$$

$$\varepsilon\frac{dy}{dt} = g_1(x,t;\varepsilon) + g_2(x,t;\varepsilon)y$$

are constructed which are valid as $\varepsilon \to 0$ throughout $0 \leq t \leq 1$.

INTRODUCTION

Consider the boundary value problem consisting of the nonlinear system

$$\left.\begin{aligned} \frac{dx}{dt} &= f(x,y,t;\varepsilon) \equiv f_1(x,t;\varepsilon) + f_2(x,t;\varepsilon)y \\ \varepsilon\frac{dy}{dt} &= g(x,y,t;\varepsilon) \equiv g_1(x,t;\varepsilon) + g_2(x,t;\varepsilon)y \end{aligned}\right\} \quad (1.1)$$

for $0 \leq t \leq 1$ and the boundary conditions

$$\left.\begin{aligned} a_1(\varepsilon)\ x(0;\varepsilon) + \varepsilon\ a_2(\varepsilon)y(0;\varepsilon) &= \alpha(\varepsilon) \\ b_1(\varepsilon)\ x(1;\varepsilon) + b_2(\varepsilon)y(1;\varepsilon) &= \beta(\varepsilon). \end{aligned}\right\} \quad (1.2)$$

Here, x and y are scalars, ε is a small, positive parameter, f and g are infinitely differentiable with respect to x and t, and f,g, the a_i, b_i, α, and β all have asymptotic power series expansions valid as $\varepsilon \to 0$.

Our object is to determine the asymptotic behavior of solutions as $\varepsilon \to 0$. Let us first suppose

(H1) that the nonlinear terminal value problem

$$\left.\begin{aligned} \frac{dX_o}{dt} &= f(X_o, -g_1(X_o,t;0)/g_2(X_o,t;0),t;0) \\ b_1(0)X_o(1) - b_2(0)g_1(X_o(1),1;0)&/g_2(X_o(1),1;0) = \beta(0) \end{aligned}\right\} \quad (1.3)$$

has a solution $X_o(t)$ on the interval $0 \leq t \leq 1$.

Then, note that the pair $(X_o(t),\ Y_o(t))$ where

$$Y_o(t) \equiv -g_1(X_o(t),t;0)/g_2(X_o(t),t;0)$$

satisfies the reduced problem consisting of the system (1.1) and the boundary condition at t = 1, both evaluated at $\varepsilon = 0$. Under appropriate additional hypotheses, we will show that the solution (x,y) of (1.1) - (1.2) will converge to (X_o,Y_o) as $\varepsilon \to 0$ away from $t = 0$. Nonuniform convergence will generally occur at $t = 0$, since we cannot expect the "boundary layer jump"

$$J = \alpha(0) - a_1(0)X_o(0)$$

to be zero.

We shall suppose further

(H2) that for some $\kappa > 0$

$$g_2(X_o(t)\underline{+}\ \theta\lambda(t),t;0) \leq -\kappa \tag{1.4}$$

and

$$|f_2(X_o(t)\underline{+}\ \theta\lambda(t),t;0)| \geq \kappa \tag{1.5}$$

throughout $0 \leq t \leq 1$ where $\lambda(t)$ is some continuous, nonvanishing function on $0 < t \leq 1$ such that $\lim_{t\to 0^+} \lambda(t) = \infty$ and $0 \leq \theta \leq 1$,

(H3) that

$$b_1(0) + b_2(0)g_x(X_o(1),Y_o(1),1;0)/g_2(X_o(1),1;0) \neq 0, \tag{1.6}$$

and

(H4) that a finite ζ exists such that

$$\int_0^\zeta [a_1(0) + a_2(0)\frac{g_2(X_o(0) + r,0;0)}{f_2(X_o(0) + r,0;0)}]dr = J \tag{1.7}$$

where J is the boundary layer jump, and

$$a_1(0) + a_2(0)\frac{g_2(X_o(0) + \zeta,0;0)}{f_2(X_o(0) + \zeta,0;0)} \neq 0. \tag{1.8}$$

Asymptotic solutions to similar two-point boundary value problems were previously obtained by Harris (1962) and Macki (1967) for small values of J. The approach here generalizes those used for second order scalar equations by O'Malley (1969) and Visik and Lyusternik (1958).

THE EXPANSION PROCEDURE

The solution will depend on a "stretched variable" or "boundary layer coordinate"

$$\tau = t/\varepsilon \tag{2.1}$$

as well as the variable t. Thus, we shall seek a solution of the form

$$\left.\begin{aligned} x(t;\varepsilon) &= X(t;\varepsilon) + m(\tau;\varepsilon) \\ y(t;\varepsilon) &= Y(t;\varepsilon) + \frac{1}{\varepsilon}\, n(\tau;\varepsilon) \end{aligned}\right\} \tag{2.2}$$

where X, Y, m, and n all have asymptotic power series expansions as $\varepsilon \to 0$. For example,

$$X(t;\varepsilon) = \sum_{j=0}^{\infty} X_j(t)\varepsilon^j.$$

(This Ansatz could be motivated by the asymptotic solution of the corresponding linear problem (cf. Wasow (1944) and O'Malley and Keller (1968)). We shall ask that the terms of m and n all tend to zero as τ tends to infinity. Thus, the outer expansion $(X(t;\varepsilon), Y(t;\varepsilon))$ will be asymptotically valid away from $t = 0$, while the "boundary layer correction" terms m and n will be significant only near $t = 0$.

The outer expansion (X, Y) must satisfy the nonlinear system and the boundary condition at $t = 1$. Substituting the expansion formally into (1.1) and this boundary condition, we have equality at $\varepsilon = 0$ since (X_o, Y_o) satisfies the reduced problem. Equating coefficients of ε^j for $j > 0$, we ask

$$\left.\begin{aligned} \frac{dX_j}{dt} &= f_x(X_o(t), Y_o(t), t;0)X_j(t) \\ &\qquad + f_2(X_o(t), t;0)Y_j(t) + p_j(t) \\ 0 &= g_x(X_o(t), Y_o(t),t;0)X_j(t) \\ &\qquad + g_2(X_o(t),t;0)Y_j(t) + q_j(t) \\ \text{and}\quad & \\ b_1(0)X_j(1) &+ b_2(0)Y_j(1) = \tilde{\beta}_j \end{aligned}\right\} \qquad (2.3)$$

where p_j, q_j and $\tilde{\beta}_j$ are determined by those X_ℓ's and Y_ℓ's with $\ell < j$. From the second equation, Y_j is determined as a linear function of X_j. The first equation, then, becomes a linear equation for X_j which becomes uniquely determined by its terminal value $X_j(1)$, known from the boundary condition. Thus, the outer expansion is uniquely obtained without any knowledge of the boundary layer correction terms.

Since the complete expansion (2.2) must satisfy (1.1) - (1.2), m and n must satisfy the nonlinear system

$$\frac{dm}{d\tau} = \varepsilon\ [(f_1(X + m,t;\varepsilon) - f_1(X,t;\varepsilon))$$

$$+ Y(t)(f_2(X + m,t;\varepsilon) - f_2(X,t;\varepsilon))] + n(\tau)f_2(X + m,t;\varepsilon)$$

$$\frac{dn}{d\tau} = \varepsilon\ [(g_1(X + m,t;\varepsilon) - g_1(X,t;\varepsilon))$$

$$+ Y(t)(g_2(X + m,t;\varepsilon) - g_2(X,t;\varepsilon))] + n(\tau)g_2(X + m,t;\varepsilon)$$

and the boundary conditions

$$a_1(\varepsilon)m(0;\varepsilon) + a_2(\varepsilon)n(0;\varepsilon) = \alpha(\varepsilon) - a_1(\varepsilon)X(0;\varepsilon) - \varepsilon a_2(\varepsilon)Y(0;\varepsilon)$$

$$b_1(\varepsilon)m(\tfrac{1}{\varepsilon};\varepsilon) + b_2(\varepsilon)n(\tfrac{1}{\varepsilon};\varepsilon) = 0.$$

(Since m and n → 0 as $\frac{1}{\varepsilon} \to \infty$, the second equation will automatically be satisfied asymptotically.) Rewriting, then, we consider a system of the form

$$\left.\begin{aligned} \frac{dm}{d\tau} &\equiv \varepsilon m(\tau)F_1(m(\tau),\varepsilon\tau;\varepsilon) + n(\tau)F_2(m(\tau),\varepsilon\tau;\varepsilon) \\ \frac{dn}{d\tau} &\equiv \varepsilon m(\tau)G_1(m(\tau),\varepsilon\tau;\varepsilon) + n(\tau)G_2(m(\tau),\varepsilon\tau;\varepsilon) \end{aligned}\right\} \qquad (2.4)$$

on the semi-infinite interval $0 \leq \tau < \infty$ where

$$a_1(\varepsilon)m(0;\varepsilon) + a_2(\varepsilon)n(0;\varepsilon) = J + 0(\varepsilon) \qquad (2.5)$$

is known and m and n → 0 as $\tau \to \infty$. Note that hypothesis (H2) implies that

$$G_2(m(\tau),\varepsilon\tau;0) \leq -\kappa < 0 \qquad (2.6)$$

and

$$|F_2(m(\tau),\varepsilon\tau;0)| \geq \kappa > 0 \qquad (2.7)$$

for m(0) arbitrary, $|m(\tau)|$ decaying, and $0 \leq \varepsilon\tau \leq 1$.

When $\varepsilon = 0$, we have

$$\frac{dm_o}{d\tau} = n_o(\tau)\ F_2(m_o,0;0)$$

$$\frac{dn_o}{d\tau} = n_o(\tau)\ G_2(m_o,0;0)$$

so

$$\frac{dn_o}{d\tau} = \frac{G_2(m_o,0;0)}{F_2(m_o,0;0)}\ \frac{dm_o}{d\tau}\ .$$

Applying the boundary conditions at infinity,

$$n_o(\tau) = \int_o^{m_o} \frac{G_2(r,0;0)}{F_2(r,0;0)}\ dr \qquad (2.8)$$

where m_o must satisfy the nonlinear equation

$$\frac{dm_o}{d\tau} = F_2(m_o,0;0) \int_o^{m_o} \frac{G_2(r,0;0)\ dr}{F_2(r,0;0)} \qquad (2.9)$$

for $0 \leq \tau < \infty$. By (2.5), the initial value $m_o(0)$ must satisfy

$$a_1(0)m_o(0) + a_2(0)\int_o^{m_o(0)} \frac{G_2(r,0;0)\,dr}{F_2(r,0;0)} = J.$$

Using hypothesis (H4), a value ζ exists so that

$$m_o(0) = \zeta. \tag{2.10}$$

By applying successive approximations to (2.9), we can obtain $m_o(\tau)$. (Generally, (2.9) can't be integrated in closed form, however.) Knowing m_o, (2.8) yields n_o. Moreover, (2.6) implies that m_o and n_o decrease monotonically to zero and

$$m_o(\tau) = 0(e^{-\kappa(1-\delta)\tau}) = n_o(\tau) \text{ as } \tau \to \infty$$

for any $\delta > 0$.

Equating coefficients of ε^j in (2.4), we ask that m_j and n_j satisfy the linear equations

$$\frac{dm_j}{d\tau} = m_j\, F_{2m}(m_o,0;0)n_o + n_j\, F_2(m_o,0;0) + P_j(\tau)$$

and

$$\frac{dn_j}{d\tau} = m_j\, G_{2m}(m_o,0;0)n_o + n_j G_2(m_o,0;0) + Q_j(\tau)$$

where P_j and Q_j are known successively. Multiplying the first equation by $-\frac{G_2(m_o,0;0)}{F_2(m_o,0;0)}$ and adding

$$\frac{dn_j}{d\tau} = \frac{d}{d\tau}\left(\frac{G_2(m_o,0;0)}{F_2(m_o,0;0)}\, m_j\right) + \tilde{Q}_j$$

with $\tilde{Q}_j$ known. Thus,

$$n_j(\tau) = \frac{G_2(m_o,0;0)}{F_2(m_o,0;0)}\, m_j(\tau) - \int_\tau^\infty \tilde{Q}_j(s)ds. \tag{2.11}$$

and

$$\frac{d}{d\tau}\left(\frac{m_j}{F_2(m_o,0;0)}\right) = m_j\frac{G_2(m_o,0;0)}{F_2(m_o,0;0)} + \tilde{P}_j(\tau),$$

where $\tilde{P}_j$ is known. Thus,

$$\begin{aligned} m_j(\tau) = F_2(m_o(\tau),0;0)\ [\ &\frac{m_j(0)}{F_2(m_o(0),0;0)} \exp\left[\int_0^\tau G_2(m_o(s),0;0)ds\right] \\ &+ \int_o^\tau \exp\left[\int_r^\tau G_2(m_o,0;0)ds\right] \tilde{P}_j(r)dr]. \end{aligned} \tag{2.12}$$

By (2.5) and (2.11), $\left(a_1(0) + a_2(0)\dfrac{G_2(m_o(0),0;0)}{F_2(m_o(0),0;0)}\right)m_j(0)$ is known successively and, by (1.8) and (2.10), this specifies $m_j(0)$ which, by (2.12) and (2.11), determines $m_j(\tau)$ and $n_j(\tau)$. Further, by induction, we can easily show that

$$m_j(\tau) = O(e^{-\kappa(1-\delta)\tau}) = n_j(\tau) \qquad \text{as } \tau \to \infty.$$

Proceeding termwise, then, the complete expansion (2.2) is obtained.

The asymptotic correctness of the formally obtained expansions is shown in O'Malley (1970b). We have:

Theorem:

For each integer $N \geq 0$, the solution of the boundary value problem (1.1) - (1.2) under hypotheses (H1) - (H4) satisfies

$$x(t;\varepsilon) = \sum_{j=0}^{N} \varepsilon^j (X_j(t) + m_j(t/\varepsilon)) + \varepsilon^{N+1} R_N(t;\varepsilon)$$

$$y(t;\varepsilon) = \frac{1}{\varepsilon} n_o(t/\varepsilon) + \sum_{j=0}^{N} \varepsilon^j (Y_j(t) + n_{j+1}(t/\varepsilon)) + \varepsilon^{N+1} S_N(t;\varepsilon)$$

where R_N and S_N are bounded for all t within $0 \leq t \leq 1$ for ε sufficiently small.

Remarks:

1. Two point problems for systems more nonlinear than (1.1) in y do not always have solutions of the form (2.2). As Coddington and Levinson (1952) observed, the example

$$\left.\begin{aligned} &\frac{dx}{dt} = y \\ &\varepsilon\frac{dy}{dt} = -y - y^3 \\ &\text{on } 0 \leq t \leq 1 \text{ with } x(0;\varepsilon) = 1,\ x(1;\varepsilon) = 0 \end{aligned}\right\}$$

has no solution for ε small. Further, the example

$$\left.\begin{aligned} &\frac{dx}{dt} = y \\ &\varepsilon\frac{dy}{dt} = -y + y^2 \\ &\text{on } 0 \leq t \leq 1 \text{ with } x(0;\varepsilon) = 1 \text{ and } x(1;\varepsilon) = 0 \end{aligned}\right\}$$

has the solution

$$\left.\begin{aligned} x(t;\varepsilon) &= -\varepsilon\, \ell n(1 + e^{-1/\varepsilon} - e^{-t/\varepsilon}) \\ y(t;\varepsilon) &= -e^{-t/\varepsilon}(1 + e^{-1/\varepsilon} - e^{-t/\varepsilon})^{-1} \end{aligned}\right\} .$$

Thus, $y(0;\varepsilon) = -e^{1/\varepsilon}$ is exponentially (instead of algebraically) large as $\varepsilon \to 0$. Finally, interior (instead of boundary layer) nonuniformities are another possibility (see O'Malley (1970a)). An example is provided by

$$\left.\begin{aligned}
&\frac{dx}{dt} = y \\
&\frac{dy}{dt} = y - y^3 \\
&\text{on } 0 \leq t \leq 1 \text{ with } x(0;\varepsilon) = 0 \text{ and } x(1;\varepsilon) = \frac{1}{2}\,.
\end{aligned}\right\}$$

The limiting solution as $\varepsilon \to 0^+$ is

$$(X_o(t),\ Y_o(t)) = \begin{cases} (0,0) & \text{for } 0 \leq t < \frac{1}{2} \\ (t - \frac{1}{2}, 1) & \text{for } \frac{1}{2} < t \leq 1 \end{cases}$$

which features nonuniform convergence at $t = \frac{1}{2}$.

2. The problem (1.1) - (1.2) when $a_2(\varepsilon) = O(\frac{1}{\varepsilon})$ is simpler than the case $a_2(\varepsilon) = O(1)$ considered. Both cases are discussed in O'Malley (1970b).

3. If g_2 were positive, the boundary layer would occur at $t = 1$ and the expansion procedure would be analogous with the stretched variable being $\tilde{\tau} = (1 - t)/\varepsilon$. If g_2 had a zero within $[0,1]$, radically different behavior could be expected. Boundary value problems for systems of the form

$$\left.\begin{aligned}
&\frac{dx}{dt} = y \\
&\varepsilon\,\frac{dy}{dt} = -\beta x + (t - \tfrac{1}{2})y
\end{aligned}\right\}$$

are discussed in Ackerberg and O'Malley (1970). Resonance phenomena occurs as β varies.

4. Examples can be found for which several values ζ exist satisfying hypothesis (H4). Each determination ζ leads to a different asymptotic solution with a common outer expansion. An example is furnished by

$$\left.\begin{aligned}
&\frac{dx}{dt} = y \\
&\varepsilon\,\frac{dy}{dt} = -\frac{1}{2}\,(1 + 3x^2)y
\end{aligned}\right\}$$

on $0 \leq t \leq 1$ with $x(0;\varepsilon) + \varepsilon y(0;\varepsilon) = 0$ and $x(1;\varepsilon) = 0$ where (H4) is satisfied for $\zeta = 0$, $\zeta = 1$, and $\zeta = -1$. These ζ values correspond to the three asymptotic solutions

$$x(t;\varepsilon) = \frac{\zeta e^{-t/2\varepsilon}}{\sqrt{(1+\zeta^2) - \zeta^2 e^{-t/\varepsilon}}}$$

$$y(t;\varepsilon) = \frac{-\zeta(1+\zeta^2)e^{-t/2\varepsilon}}{2\varepsilon\sqrt{((1+\zeta^2)-\zeta^2e^{-t/\varepsilon})^3}} \quad .$$

Likewise, the boundary condition for (1.3) might allow several different determinations of $X_o(1)$ and $X_o(t)$.

5. If $f_2(x,t;\varepsilon) \equiv 0$, we can obtain analogous results (with $m(\tau;\varepsilon) \equiv 0$) by replacing (H2) and (H4) by

(H2)' for some $\kappa > 0$, $g_2(X_o(t),t;0) \leq -\kappa$ for $0 \leq t \leq 1$ and

(H4)' $a_2(0) \neq 0$.

REFERENCES

1. R. C. Ackerberg and R. E. O'Malley, Jr., "Boundary layer problems exhibiting resonance," *Studies in Appl. Math*. 49 (1970).

2. E. A. Coddington and N. Levinson, "A boundary value problem for a nonlinear differential equation with a small parameter," *Proc. Amer. Math. Soc*. 3 (1952), pp. 73-81.

3. W. A. Harris, Jr., "Singular perturbations of a boundary value problem for a nonlinear system of differential equations," *Duke Math. J*. 29 (1962), pp. 429-445.

4. J. W. Macki, "Singular perturbations of a boundary value problem for a system of nonlinear ordinary differential equations," *Arch. Rational Mech. Anal*. 24 (1967), pp. 219-232.

5. R. E. O'Malley, Jr., "On a boundary value problem for a nonlinear differential equation with a small parameter," *SIAM J. Appl. Math*. 17 (1969), pp. 569-581.

6. R. E. O'Malley, Jr., "On nonlinear singular perturbation problems with interior nonuniformities," *J. Math. Mech*. 19 (1970), pp. 1103-1112.

7. R.E. O'Malley, Jr., "Singular perturbations of a boundary value problem for a system of nonlinear differential equations," J. Diff. Eq. 8 (1970).

8. R. E. O'Malley, Jr., and J. B. Keller, "Loss of boundary conditions in the asymptotic solution of linear ordinary differential equations, II, boundary value problems," *Comm. Pure Appl. Math*. 21 (1968), pp. 263-270.

9. M. I. Visik and L. A. Lyusternik, "On the asymptotic behavior of the solutions of boundary problems for quasi-linear differential equations,"

Dokl. Akad. Nauk SSSR 121 (1958), pp. 778-781.

10. W. Wasow, "On the asymptotic solution of boundary value problems for ordinary differential equations containing a parameter," J. Math. Phys. 23 (1944), pp. 173-183.

This research was supported by the Air Force Office of Scientific Research under contract number AFOSR-537-67.

ON A CLASS OF SINGULARLY PERTURBED TWO POINT BOUNDARY VALUE PROBLEMS

CHARLES R. HADLOCK
Department of Mathematics
Amherst College
Amherst, Massachusetts

ABSTRACT

This paper investigates the limiting behavior, as a small positive parameter μ approaches 0, of a nonlinear two point boundary value problem containing μ multiplying some of the derivatives.

INTRODUCTION

Two point boundary value problems arise in optimal control theory in the formulation of necessary conditions for optimality [1]. The difficulty of solving such problems, especially for high order nonlinear systems, motivates the study of methods for treating the original system by means of computations restricted to one or several lower order systems derived from it [2]. This paper investigates a class of two point boundary value problems for which a "singular perturbation" technique is applicable [3]. The type of boundary conditions treated herein typically arise in fixed time, fixed initial point, free endpoint problems, although the results can be extended to other problems also. The methods used are largely an extension of those of [4, 5], wherein the central subject was the behavior of the stable initial manifolds of certain differential equations.

A CONVERGENCE RESULT

In this section sufficient conditions are given for the existence of a solution to the full system

$$\dot{x} = f(x,v,y,w,t,\mu) \tag{1a}$$

$$\dot{v} = p(x,v,y,w,t,\mu) \tag{1b}$$

$$\mu\dot{y} = g(x,v,y,w,t,\mu) \tag{1c}$$

$$\mu\dot{w} = q(x,v,y,w,t,\mu) \tag{1d}$$

$$x(t^o) = x^o \tag{2a}$$

$$v(T) = v^T \tag{2b}$$

$$y(t^o) = y^o \tag{2c}$$

$$w(T) = w^T \tag{2d}$$

on $[t^o, T]$, where "$\dot{}$" denotes $\frac{d}{dt}$, μ is a small positive parameter, and x, v, y, w have respective dimensions n_1, n_2, n_3, n_4; and for convergence of this solution as $\mu \to 0$ to a known, isolated solution $\bar{x}(t), \bar{v}(t), \bar{y}(t), \bar{w}(t)$ of the degenerate system

$$\dot{x} = f(x,v,y,w,t,0) \tag{3a}$$

$$\dot{v} = p(x,v,y,w,t,0) \tag{3b}$$

$$0 = g(x,v,y,w,t,0) \tag{3c}$$

$$0 = q(x,v,y,w,t,0) \tag{3d}$$

$$x(t^o) = x^o \tag{4a}$$

$$v(T) = v^T \tag{4b}$$

gotten by setting $\mu = 0$ in the original. Since the number of derivatives has been reduced in going from (1) to (3), it is natural only to specify auxiliary conditions on x and v for the degenerate system. The perturbation of the parameter from a positive value to 0 changes the nature of the system from a differential equation to a combination of a differential equation and a functional equation, necessitating in the process the sacrifice of the auxiliary conditions on y and w. Hence it is called a singular perturbation.

It is convenient to make the following abbreviations:

$$X \equiv \begin{pmatrix} x \\ v \end{pmatrix} \qquad \bar{X} \equiv \begin{pmatrix} \bar{x} \\ \bar{v} \end{pmatrix}$$

$$Y \equiv \begin{pmatrix} y \\ w \end{pmatrix} \qquad \bar{Y} \equiv \begin{pmatrix} \bar{y} \\ \bar{w} \end{pmatrix}$$

$$F \equiv \begin{pmatrix} f \\ p \end{pmatrix} \qquad G \equiv \begin{pmatrix} g \\ q \end{pmatrix} .$$

The original system equations may thus be abbreviated:

$$\dot{X} = F(X,Y,t,\mu)$$

$$\mu\dot{Y} = G(X,Y,t,\mu)$$

and the degenerate system equations may be abbreviated:

$$\dot{X} = F(X,Y,t,0)$$

$$0 = G(X,Y,t,0) .$$

We define the matrix derivatives $F_X(t) \equiv F_X(\bar{X}(t), \bar{Y}(t), t, 0)$ with similar definitions for $F_Y(t), G_X(t)$ and $G_Y(t)$. The following is the main result of this section:

Theorem 1: Assume the following hypotheses:

H1) There exists a solution $\bar{X}(t)$, $\bar{Y}(t)$ to the degenerate system (3) and (4).

H2) *There exists a* $\delta_o > 0$ *such that in the region* D_o *of* (X,Y,t,μ) *space defined by*

$$\begin{aligned} &|X - X(t)| < \delta_o \\ &|Y - Y(t)| < \delta_o \\ &t^o \leq t \leq T \\ &0 < \mu < \delta_o \end{aligned}$$

the functions F *and* G *and the derivatives* $F_X, F_Y, G_X, G_Y, F_\mu, G_\mu$ *exist and are continuous.*

H3) *There exists a non-singular, continuously differentiable square matrix* $P(t)$, *of dimension* $n_3 + n_4$, *such that*

$$P^{-1}(t)G_Y(t)P(t) = \begin{pmatrix} C(t) & 0 \\ 0 & D(t) \end{pmatrix}$$

where $C(t)$ *is an* $n_3 \times n_3$ *square matrix each of whose eigenvalues, for each* t, *has a negative real part, and* $D(t)$ *is an* $n_4 \times n_4$ *square matrix each of whose eigenvalues, for each* t, *has a positive real part.*

H4) $G_Y^{-1}(t)G_X(t) \in C^1[t^o,T]$.

H5) *The matrix* $P(t)$, *when partitioned into the form*

$$\begin{pmatrix} P_1(t) & P_2(t) \\ P_3(t) & P_4(t) \end{pmatrix}$$

where $P_1(t)$ *is* $n_3 \times n_3$ *and* $P_4(t)$ *is* $n_4 \times n_4$, *satisfies:*

$$\begin{aligned} &P_1(t^o) \text{ is a non-singular matrix} \\ &P_4(T) \text{ is a non-singular matrix.} \end{aligned}$$

H6) *The fundamental matrix* $\Phi(T,t^o)$ *of* $\dot{Z} = (F_X(t) - F_Y(t)G_Y^{-1}(t)G_X(t))Z$, *when partitioned into the same size blocks as* $P(t)$ *in H5, satisfies: at least one of the two diagonal blocks is a non-singular matrix.*

Then there exists a $\gamma^* > 0$ *such that whenever* $0<\mu\leq\gamma^*$ *and* $|y^o - \bar{y}(t^o)|\leq\gamma^*$ *and* $|w^T - \bar{w}(T)| \leq \gamma^*$, *then a solution to equations (1) and (2) exists on* $[t^o,T]$. *Moreover, there exist positive constants* K *and* δ, *independent of* μ, *and a function* $\omega(\mu) \in C^0[0,\gamma^*]$, $\omega(0) = 0$, *such that this solution satisfies:*

$$|x(t,\mu) - \bar{x}(t)| \leq \omega(\mu) \qquad (5a)$$

$$|v(t,\mu) - \bar{v}(t)| \leq \omega(\mu) \qquad (5b)$$

$$|y(t,\mu) - \bar{y}(t)| \le Ke^{-\sigma\frac{t-t^{o}}{2\mu}} + \omega(\mu) + Ke^{-\sigma\frac{T-t}{2\mu}} \quad (5c)$$

$$|w(t,\mu) - \bar{w}(t)| \le Ke^{-\sigma\frac{t-t^{o}}{2\mu}} + \omega(\mu) + Ke^{-\sigma\frac{T-t}{2\mu}} \quad (5d)$$

From (5), the solution $X(t,\mu)$ converges to $\bar{X}(t)$, as $\mu \to 0^+$, uniformly on the entire closed interval, whereas $Y(t,\mu)$ converges to $\bar{Y}(t)$ on the open interval (t^{o},T) and uniformly on closed subsets of (t^{o},T). The lack of convergence of $Y(t,\mu)$ to $\bar{Y}(t)$ at the endpoints is called boundary layer behavior; it results from the fact that y^{o} and $\bar{y}(t^{o})$ will in general be different, as will w^{T} and $\bar{w}(T)$. In the above inequalities, the possibility of such behavior is indicated by the exponential terms. Note that from inequalities (5c) and (5d) it is clear that there may be a boundary layer in both y and w at each end of the interval. One might expect, since only an initial condition was lost on the y variable, that the third term on the right of (5c) would be absent, and for similar reasons the first term on the right of (5d). Their inclusion results from the coupling of the equations. The proof of Theorem 1 is accomplished with the aid of several lemmas, summarized below:

Lemma 1: Let $A(t) = F_X(t) - F_Y(t)G_Y^{-1}(t)G_X(t)$. *Then the matrix differential equation*

$$\dot{Q} = A(t)Q$$

has a square $n_3 + n_4$ *dimensional non-singular solution* $Q(t)$ *on* $[t^{o},T]$ *which, when partitioned into the form*

$$Q(t) = \begin{pmatrix} Q_1(t) & Q_2(t) \\ Q_3(t) & Q_4(t) \end{pmatrix}$$

satisfies $Q_2(t^{o}) = 0$ *and* $Q_3(T) = 0$.

This lemma is an immediate consequence of H6 and the fact that solutions of a linear system are independent on an interval if and only if they are independent at some point in that interval.

Lemma 2: Let $Q(t)$ *be a matrix resulting from the application of Lemma 1 . Then the change of variables*

$$\begin{pmatrix} x - \bar{x}(t) \\ v - \bar{v}(t) \end{pmatrix} = Q(t)\begin{pmatrix} \pi \\ \rho \end{pmatrix} \quad (6a)$$

$$\begin{pmatrix} y - \bar{y}(t) \\ w - \bar{w}(t) \end{pmatrix} = P(t) \begin{pmatrix} \eta \\ \phi \end{pmatrix} - G_Y^{-1}(t) G_X(t) Q(t) \begin{pmatrix} \pi \\ \rho \end{pmatrix} \tag{6b}$$

transforms the original system (1) into the following form:

$$\dot{\pi} = B_1(t)\eta + B_2(t)\phi + S_1(\pi,\rho,\eta,\phi,t,\mu) \tag{7a}$$

$$\dot{\rho} = B_3(t)\eta + B_4(t)\phi + S_2(\pi,\rho,\eta,\phi,t,\mu) \tag{7b}$$

$$\mu\dot{\eta} = C(t)\eta + S_3(\pi,\rho,\eta,\phi,t,\mu) \tag{7c}$$

$$\mu\dot{\phi} = D(t)\phi + S_4(\pi,\rho,\eta,\phi,t,\mu) \tag{7d}$$

The new variables π,ρ,η,ϕ *are of respective dimensions* n_1,n_2,n_3,n_4. *The vector functions* S_i *each satisfy the following 2 properties:*

A) For each $\delta>0$ $\exists$ $\gamma(\delta)>0$ *s.t. for* $|\pi_1|,|\pi_2|$, *etc. and* $\mu<\gamma$

$$|S_i(\pi_1,\rho_1,\eta_1,\phi_1,t,\mu) - S_i(\pi_2,\rho_2,\eta_2,\phi_2,t\mu)| \leq \delta(|\pi_1-\pi_2| + |\rho_1-\rho_2| + |\eta_1-\eta_2| + |\Phi_1-\Phi_2|)$$

B) There exists a function $\omega(\mu) \in C^0[0,\mu^o]$ *for some* $\mu^o > 0$, $\omega(0) = 0$, *such that* $|S_i(0,0,0,0,t,\mu| \leq \omega(\mu)$.

This lemma is a convenient linearization of the system around the degenerate solution, properties A and B, consequences of the mean value theorem, being written in a form useful to the subsequent development.

Lemma 3: *There exist positive constants* K_1 *and* γ_1 *and a function* $\omega(\mu) \in C^0[0,\gamma_1], \omega(0)=0$, *s.t. if* $|a|,|b|,|c|,|d|,\mu$ *are each* $\leq\gamma_1$, *(of course,* $\mu>0$*), then system (7) has a solution satisfying* $\pi(t^o) = a$, $\rho(T) = b$, $\eta(t^o) = C$, $\phi(T) = d$, *and*

$$|\pi(t,\mu)| \leq K_1[|a| + |b| + \mu|c| + \mu|d| + \omega(\mu)] \tag{8a}$$

$$|\rho(t,\mu)| \leq K_1[|a| + |b| + \mu|c| + \mu|d| + \omega(\mu)] \tag{8b}$$

$$|\eta(t,\mu)| \leq K_1[|a| + |b| + \mu|c| + |c|e^{-\sigma\frac{t-t^o}{2\mu}} + \mu|d| + |d|e^{-\sigma\frac{T-t}{2\mu}} + \omega(\mu)] \tag{8c}$$

$$|\phi(t,\mu)| \leq K_1[|a| + |b| + \mu|c| + |c|e^{-\sigma\frac{t-t^o}{2\mu}} + \mu|d| + |d|e^{-\sigma\frac{T-t}{2\mu}} + \omega(\mu)]. \tag{8d}$$

The basic outline of the proof is the following. We show that there exists a $\delta > 0$ such that if $|a|,|b|,|c|,|d|,\mu$ are each $\leq \gamma(\delta)$, as chosen by Property A of Lemma 2, then the method of successive approximations applied to the integral equations resulting from the application of the variation of parameters formula to (7) gives a solution. The bounds on the differences of successive iterates sum to expressions of the form of the right sides of (8).

Lemma 4: *Define* $\bar{\pi}(t) \equiv \pi(a,b,c_1,d_1,t,\mu)$, $\hat{\pi}(t) \equiv \pi(a,b,c_2,d_2,t,\mu)$, *and similarly for the other functions. These functions are understood to be those derived in the proof of Lemma 3 and so we assume* $|a|,|b|,|c_1|,|c_2|,|d_1|,|d_2|$, *and* μ *are each* $\leq \gamma_1$.

Then there exist positive constants K_2 *and* δ *such that for each* δ, $0 \leq \delta \leq \delta_o$, *whenever* $|a|,|b|,|c_1|,|c_2|,|d_1|,|d_2|,\mu$ *are each* $\leq \gamma(\delta)$, $\gamma(\delta)$ *being the same as in Property A,*

$$|\bar{\pi}(t) - \hat{\pi}(t)| \leq K_2(\mu|c_1-c_2| + \mu|d_1-d_2|) \tag{9a}$$

$$|\bar{\rho}(t) - \hat{\rho}(t)| \leq K_2(\mu|c_1-c_2| + \mu|d_1-d_2|) \tag{9b}$$

$$\begin{aligned}|\bar{\eta}(t) - \hat{\eta}(t)| \leq\ & K_2(\mu|c_1-c_2| + \mu|d_1-d_2|) \\ & + |c_1-c_2|e^{-\sigma\frac{t-t^o}{2\mu}} + \delta|d_1-d_2|e^{-\sigma\frac{T-t}{2\mu}})\end{aligned} \tag{9c}$$

$$\begin{aligned}|\bar{\phi}(t) - \hat{\phi}(t)| \leq\ & K_2(\mu|c_1-c_2| + \mu|d_1-d_2| \\ & + \delta|c_1-c_2|e^{-\sigma\frac{t-t^o}{2\mu}} + |d_1-d_2|e^{-\sigma\frac{T-t}{2\mu}}).\end{aligned} \tag{9d}$$

This lemma, in particular the δ factor in one term of each (9c) and (9d), is the key to the proof of Theorem 1. The proof of the lemma is based on a comparison of the i^{th} iterates in the successive approximations leading to $\bar{\pi}$ and $\hat{\pi}$, etc.

Lemma 5: *Define* $\pi^*(t) \equiv \pi(0,0,0,0,t,\mu)$ *and similarly* ρ^*, η^*, *and* ϕ^*. *For any* $\varepsilon > 0$ *there exists a* $\mu_1(\varepsilon)$ *such that for* $\mu \leq \mu_1$ $|\pi^*(t)|,|\rho^*(t)|,|\eta^*(t)|,|\phi^*(t)|$ *will each be* $\leq \varepsilon$.

Lemma 6: *For any* $\varepsilon>0$ *there exists a positive* $\gamma_2(\varepsilon) \leq \gamma(\delta_o)$ $|a|,|b|,|c_1|,|c_2|,|d_1|,|d_2|$, *each* $\leq \gamma_2$

$$|\bar{\pi}(T) - \hat{\pi}(T)| \leq \varepsilon(|c_1-c_2| + |d_1-d_2|) \tag{10a}$$

$$|\bar{\rho}(t^o) - \hat{\rho}(t^o)|\ \underline{\;}\ \varepsilon(|c_1-c_2| + |d_1-d_2|) \tag{10b}$$

$$|\bar{\eta}(T) - \hat{\eta}(T)|\ \underline{\;}\ \varepsilon(|c_1-c_2| + |d_1-d_2|) \tag{10c}$$

$$|\bar{\phi}(t^o) - \hat{\phi}(t^o)| \leq \varepsilon(|c_1-c_2| + |d_1-d_2|) \tag{10d}$$

These two lemmas are straightforward consequences of Lemmas 3 and 4 respectively.

We are now in a position to outline the proof of Theorem 1. The basic problem is to find values of a, b, c, and d that give solutions of (7) which in terms of the original variables satisfy the required boundary conditions. It follows from the definition of $Q(t)$ that the correct values of a and b are 0. The problem is to demonstrate the existence of values for c and d satisfying (recall (6)):

$$y^o-\bar{y}(t^o) = P_1(t^o)\eta(0,0,c,d,t^o,\mu) + P_2(t^o)\phi(0,0,c,d,t^o,\mu)$$

$$-R_1(t^o)\pi(0,0,c,d,t^o,\mu) - R_2(t^o)\rho(0,0,c,d,t^o,\mu) \qquad (11a)$$

$$w^T-w(T) = P_3(T)\eta(0,0,c,d,T,\mu) + P_4(T)\phi(0,0,c,d,T,\mu)$$

$$-R_3(T)\eta(0,0,c,d,T,\mu) - R_4(T)\phi(0,0,c,d,T,\mu) \qquad (11b)$$

where we have made the abbreviation $R(t)\equiv G_y^{-1}(t)G_x(t)Q(t)$. These are clearly equivalent to

$$y^o-\bar{y}(t^o) = P_1(t^o)c + P_2(t^o)\phi(0,0,c,d,t^o,\mu)$$

$$- R_2(t^o)\rho(0,0,c,d,t^o,\mu)$$

$$w^T-\bar{w}(T) = P_3(T)\eta(0,0,c,d,T,\mu) + P_4(T)d - R_3(T)\pi(0,0,c,d,T,\mu)$$

In view of H5, these can be rewritten

$$c = P_1^{-1}(t^o)P_2(t^o)\phi(0,0,c,d,t^o,\mu) +$$

$$+ P_1^{-1}(t^o)R_2(t^o)\rho(0,0,c,d,t^o,\mu) + \bar{y}(t^o) - y^o$$

$$d = -P_4^{-1}(T)P_3(T)\eta(0,0,c,d,T,\mu) + P_4^{-1}(T)R_3(T)\pi(0,0,c,d,T,\mu)$$

$$+ \bar{w}(T) - w^T$$

By defining the function, for each fixed μ,

$$H(c,d) \equiv \begin{pmatrix} H_1(c,d) \\ \\ H_2(c,d) \end{pmatrix} \equiv \begin{pmatrix} -P_1^{-1}(t^o)P_2(t^o)\phi(0,0,c,d,t^o,\mu) \\ + P_1^{-1}(t^o)R_2(t^o)\rho(0,0,c,d,t^o,\mu) \\ + \bar{y}(t^o) - y^o \\ -P_4^{-1}(T)P_3(T)\eta(0,0,c,d,T,\mu) \\ + P_4^{-1}(T)\pi(0,0,c,d,T,\mu) \\ + \bar{w}(T) - w^T \end{pmatrix}$$

the solution of (11) consists in finding a suitable fixed point of H. This can be accomplished by using Lemmas 5 and 6 to show that for $|c|,|d|$ and μ sufficiently small, H is a contraction mapping. Inequalities (5) then follow immediately from (9).

CONSTRUCTION OF AN EXPANSION

Theorem 1 gives conditions under which it is reasonable to try to approximate the solution of the full problem by that of the degenerate problem, which in general is easier to solve. This initial approximation may be improved upon by constructing a parametric expansion to the solution of the full problem. For the problem treated by Theorem 1, the details of this method may be found in [6].

Let $z(t,\mu)$ denote the composite vector made up of $x(t,\mu)$, $v(t,\mu)$, $y(t,\mu)$, and $w(t,\mu)$. One seeks an nth order asymptotic representation of $z(t,\mu)$ in the form

$$z(t,\mu) \sim \sum_{i=0}^{n} L_i z(\tau_L)\mu^i + \sum_{i=0}^{n} \tilde{z}_i(t)\mu^i + \sum_{i=0}^{n} R_i z(\tau_R)\mu^i \tag{12}$$

where $\tau_L \equiv \frac{t-t^o}{\mu}$ and $\tau_R \equiv \frac{T-t}{\mu}$. One substitutes the right side of (12) for $z(t,\mu)$ in equations (1), equating separately corresponding coefficients of τ_L, t, and τ_R. The result is a sequence of recursive differential and functional equations for the coefficient functions $L_i z$, $\tilde{z}_i$, and $R_i z$. The auxiliary conditions for these equations are determined by (2) and the assumption that the $L_i z$ and $R_i z$ decay exponentially as their arguments increase. With these auxiliary conditions there result linear two point boundary value problems for the coefficient functions of order one and higher. The zeroth order term $\tilde{z}_0(t)$ is the degenerate solution $\bar{z}(t)$. Although $L_0 z$ and $R_0 z$ may be sought via a two point boundary value problem, certain computational complications can be avoided when $L_0 w(0)$ and $R_0 y(0)$ can be calculated directly, for then $L_0 z$ and $R_0 z$ can be found by solving an initial value problem. From (5a) and (5b) it is clear that $L_0 x$, $L_0 v$, $R_0 x$, $R_0 v$ are all identically 0; and so, in view of (6), $L_0 w(0)$ and $R_0 y(0)$ can be found if the limit as $\mu \to 0$ of $y(T,\mu)$ and $w(t^o,\mu)$ can be calculated. For a certain class of systems the following theorem is applicable.

*Theorem 2**: In the context of Theorem 1, and if, in addition, $G(X,Y,t,0)$ is linear in Y, then*

$$\lim_{\mu \to 0} y(T,\mu) = \bar{y}(T) + P_2(T)P_4^{-1}(T)(w^T - \bar{w}(T))$$

$$\lim_{\mu \to 0} y(t^o,\mu) = \bar{w}(t^o) + P_3(t^o)P_1^{-1}(t^o)(y^o - \bar{y}(t^o))$$

The proof of this theorem in [6] is for a slightly more general class of systems.

The approach of this paper has been applied to the mathematical model of the control of a strip winding process in a rolling mill plant [7].

REFERENCES

1] Athans, M. and P. L. Falb, Optimal Control, McGraw-Hill, New York, 1966.

2] Kokotović, P. and P. Sannuti, "Singular Perturbation Method for Reducing the Model Order in Optimal Control Design," IEEE Trans. on Automatic Control, Vol. 13 (1968), pp. 377-384.

3] Wasow, W., Asymptotic Expansions for Ordinary Differential Equations, Interscience, New York, 1965, Chapter 10.

4] Levin, J. J., "Singular Perturbations of Nonlinear Systems of Differential Equations Related to Conditional Stability," Duke Math. J., Vol. 23 (1956), pp. 609-620.

5] Levin, J. J., "The Asymptotic Behavior of the Stable Initial Manifolds of a System of Nonlinear Differential Equations," AMS Trans., Vol. 85 (1957), pp. 357-368.

6] Hadlock, C., "Singular Perturbation of a Class of Two Point Boundary Value Problems Arising in Optimal Control," Report R-481, Coordinated Science Laboratory, University of Illinois, Urbana, 1970.

7]Hadlock, C., M. Jamshidi, and P. Kokotović, "Near Optimum Design of Three Time Scale Systems," Proc. 4th Annual Princeton Conference on Information Science and Systems, Princeton, N.J., March, 1970.

ACKNOWLEDGMENT

This work was supported in part by the Joint Services Electronics Program (U. S. Army, U. S. Navy, and U. S. Air Force) under Contract DAAB-07-67-C-0199, in part by the U. S. Air Force under Grant AFOSR-68-1579C and in part by the National Science Foundation under Grant GK-3893, Coordinated Science Laboratory, University of Illinois.

SINGULARLY PERTURBED BOUNDARY VALUE PROBLEMS WITH INTERIOR TURNING POINTS

W. A. HARRIS, JR.*
Department of Mathematics
University of Southern California
Los Angeles, California 90007

ABSTRACT

In this note we discuss two physically motivated problems of singular perturbation type. Both problems represent the simplest cases in a certain sense, but the techniques utilized represent current mathematical research.

1. Introduction

In this note we will discuss two physically motivated problems of singular perturbation type - a problem consisting of a differential equation involving a small parameter ϵ and some subsidiary conditions in which the order of the differential equation is lower for $\epsilon = 0$ than for $\epsilon \neq 0$. In general, the object of perturbation theory is to solve a problem approximately for small ϵ when the solution is known for $\epsilon = 0$. If the differential equation and subsidiary conditions depend on ϵ in a sufficiently regular manner, standard continuity theorems guarantee the desired result. However, the simplest examples show that the situation is more complicated for singular perturbation problems.

Both problems that we will discuss are in a certain sense the simplest cases; but the validity of the techniques utilized in their solutions represents current mathematical research. We hope that our presentation will serve a an imputus to further research in extending the validity of these techniques.

2. Two Point Boundary Value Problems

Let us consider the real boundary value problem

$$\begin{gathered} \epsilon y'' + a(x,\epsilon)y' + b(x,\epsilon)y = 0 \\ y(-1),\epsilon) = \alpha, \ y\ (1,\epsilon) = \beta \end{gathered} \tag{2.1}$$

on the interval $-1 \leq x \leq 1$, where ϵ is a small positive parameter and $a(x,\epsilon)$, $b(x,\epsilon)$ are sufficient regular functions having a symptotic expansions in nonnegative power of ϵ as $\epsilon \to 0$. The problem at hand

---

* Supported in part by the United States Army under Contract DA - ARO - D - 31 - 124 - 71 - G14.

is to determine under what conditions the solution $y(x, \epsilon)$ of the boundary value problem (2.1) will converge to a solution of the reduced differential equation

$$a(x, 0) z' + b(x, 0) z = 0$$

and to determine the appropriate reduced boundary condition.

Boundary value problems for linear differential equations have been studied extensively, see for example Wasow [9,10] and Harris [2]. In particular, the solution of the boundary value problem

$$\epsilon y'' + a(x, \epsilon) y' + b(x, \epsilon) y = 0$$
$$y(-1, \epsilon) = \alpha, \quad y(1, \epsilon) = \beta$$

will converge to a solution of the reduced boundary value problem

$$a(x, 0) z' + b(x, 0) z = 0$$
$$z(-1) = \alpha$$

for $a(x, \epsilon) < 0$ as $\epsilon \rightarrow 0$ except at $x = 1$, and

$$a(x, 0) z' + b(x, 0) z = 0$$
$$z(1) = \beta$$

for $a(x, \epsilon) > 0$ as $\epsilon \rightarrow 0$ except at $x = -1$. In general, boundary layer behavior occurs at the excepted endpoint.

Following convention, we say that the differential equation

$$\epsilon y'' + a(x, \epsilon) y' + b(x, \epsilon) y = 0$$

has a turning point at x_o if $a(x_o, 0) = 0$ but $a(x, 0) \not\equiv 0$, or if $b(x_o, 0) = 0$, $b(x, 0) \not\equiv 0$, $a(x, \epsilon) \equiv 0$. We note that it is always possible to transform this differential equation into one for which the first derviative term is missing and most classifications and studies begin with the differential equation in this special form.

The boundary value problem

$$\epsilon y'' + 4cxy' - 4cdy = 0$$
$$y(-1, \epsilon) = \alpha, \quad y(1, \epsilon) = \beta \tag{2.2}$$

with constants $c > 0$, d represents a simple example of a singularly perturbed boundary value problem with an interior turning point at $x = 0$. In view of the results for boundary value problems without turning points, we might expect the corresponding reduced boundary value problem to be

$$xz_1' - dz_1 = 0, \quad z_1(-1) = y(-1) = \alpha \quad \text{on } [-1, 0),$$
$$xz_2' - dz_2 = 0, \quad z_2(1) = y(1) = \beta \quad \text{on } (0, 1] \tag{2.3}$$

with behavior at the turning point $x = 0$ more complicated and depending on the sign of d.

Indeed we can show that this is essentially the case since we can solve the differential equation

$$\varepsilon y'' + 4cxy' - 4cdy = 0 \tag{2.4}$$

in terms of well known special functions - Weber functions or parabolic cylinder functions.

The change of dependent and independent variable

$$y(x) = z(x) \exp \left\{ -\frac{cx^2}{\varepsilon} \right\}, \quad t = \sqrt{\frac{4c}{\varepsilon}}\, x$$

transform the differential equation (2.4) into

$$\frac{d^2 z}{dt^2} - (1/2 + d + t^2/4) z = 0 \tag{2.5}$$

Recalling that the parabolic cylinder functions $D_n(t)$ and $D_{-(n+1)}(it)$ are linearly independent solutions of the equation

$$\frac{d^2 w}{dt^2} + (n + 1/2 - t^2/4)\, w = 0,$$

the solution of the boundary value problem (2.2) can be written

$$y(x, \varepsilon) = e^{-\frac{cx^2}{\varepsilon}} \left\{ c_1 D_{-1-d}\left(\sqrt{\frac{4c}{\varepsilon}}\, x\right) + c_2 D_d\left(i \sqrt{\frac{4c}{\varepsilon}}\, x\right) \right\}$$

where c_1 and c_2 are determined by the pair of linear equations corresponding to the boundary conditions

$$y(-1, \varepsilon) = \alpha, \quad y(1, \varepsilon) = \beta .$$

In order to evaluate the limit of the solution $y(x, \varepsilon)$ as $\varepsilon \to 0$, we note that the parabolic cylinder function $D_n(z)$ is an entire function having the asymptotic expansions

$$D_n(z) \sim e^{-\frac{z^2}{4}} z^n (1 + o(1)) \quad \text{as } |z| \to \infty, \ |\arg z| < \frac{3\pi}{4}$$

$$D_n(z) \sim e^{-\frac{z^2}{4}} z^n (1 + o(1)) - \frac{\sqrt{2\pi}}{\Gamma(-n)} \frac{e^{n\pi i}\, e^{z^2/4}}{z^{n+1}} (1 + o(1)) \tag{2.6}$$

$$\text{as } |z| \to \infty \quad |\arg z - \pi| < \frac{3\pi}{4} .$$

The behavior of $D_n(x)$ for x negative, $|x| \to \infty$ is substantially different depending upon whether n is a nonnegative integer or not.

Utilizing the asymptotic exansions (2.6) we can show that the solution $y(x, \epsilon)$ of the boundary value problem (2.2) has the asymptotic form, if d is not a negative integer,

$$y(x, \epsilon) \sim z_2\,(x) \text{ for } x > 0$$

$$y(x, \epsilon) \sim z_1\,(x) \text{ for } x < 0$$

$$y(0, \epsilon) = 0\,(\epsilon^{d/2})$$

where z_1, z_2 are solutions of the reduced boundary value problem (2.3). For a discussion of the case $c > 0$, d a negative integer and the case $c < 0$ and d not a nonnegative integer and a nonnegative integer see O'Malley [6] from which these results are taken.

3. A more general boundary value problem.

One would expect that the more general boundary value problem

$$\epsilon y'' + xa(x, \epsilon)\, y' + b(x, \epsilon)\, y = 0$$
$$(3.1)$$
$$y(-1, \epsilon) = \alpha\,, \; y\,(1, \epsilon) = \beta$$

with $a(x, o) \neq 0$, $a\,(x, \epsilon\,)$ and $b(x, \epsilon)$ sufficiently regular with asymptotic expansions in nonnegative powers of ϵ should exhibit the same qualitative behavior as the corresponding boundary value problem discussed in Section 2 with a and b constants. In order to show that this is the case, we need detailed asymptotic information about an appropriate set of linearly independent solutions of the corresponding differential equation analogous to the information contained in formula (2.6). This formula exhibits one of the basic properties of asymptotic representation of solutions of differential equations, namely the Stokes phenomenon, that the asymptotic form of a given solution of the differential equation (2.5) will vary in the neighborhood of the irregular singular point $t = \infty$. Indeed, there are essentially two formal series, one of the form $w_1 = e^{-z^2/4} z^n\,(1 + o(1)\,)$, and one of the form $w_2 = e^{z^2/4} z^{-(n+1)}(1+o(1))$. The determination of a coefficient $c(z)$ such that $D_n(z) \sim w_1 + c(z)\, w_2$ as indicated in formula (2.6) is called a connection problem and the coefficient c is called a connection coefficient or Stokes' multiplier. We note that the solution of this connection problem for the Weber functions was fundamental is solving the boundary value problem (2.2).

There are essentially two general methods for obtaining the required linearly independent set of solutions with appropriate asymptotic representations and Stokes' multipliers. The first method is that of related equation as typified by R. E. Langer [4] et al, and the second method is that of analytic simplification as typified by Y. Sibuya, [7] et al.

In the method of related equations we choose a basic equation such as the Airy equation or the Weber equation for which we have complete knowledge - i.e. well known special functions. We then choose a related equation, a transformed basic equation, to compare with the given differential equation which is to approximate the give differential equation in a suitable fashion so that solution forms of the given differential equation can be deduced from the corresponding known forms of the related equation. To be useful in the present context we would insist that this information include the Stokes' multipliers. Great ingenuity and analytic technique is needed in this method which often seems ad hoc.

In the method of analytic simplification we begin with the original differential equation and attempt to simplify the equation formally. If we are ingenious, we will obtain an equation which should represent the basic problem. We then give analytic meaning to the formal simplification and indeed if we have achieved a uniform simplification, we have reduced the problem to a basic one for further study. Great progress has been made using the analytic simplification method but unfortunately the problem of uniform simplification is essentially equivalent to solving the connection problem for the original equation in terms of Stokes' multipliers for the basic equation. Indeed, only for problems which can be formally simplified to yield a basic problem which is the Airy equation or the Weber equation has the uniform simplification been established, see Lee [5]. O'Malley [6] has shown that the boundary value problem (3.1) can be formally simplified to a boundary value problem of the type (2.2) and used Lee's results on uniform simplification to obtain analogous results.

The uniform simplification of problems reducible to the Airy or Weber equation give hope that the solution of the connection problem can be inferred from the solution of the connection problem for the corresponding basic problem - at least for second order differential equation without secondary turning points in the sense of Sibuya and Iwano [3]. Indeed further consideration of the fundamental problem is warrented since it is frequently tacitly assumed in the study of applied problems, for example in studies of the Orr - Sommerfield equation.

4. Eigenvalue problems.

Let us consider the eigenvalue problem

$$y''(x) - \lambda^2 p(x) y(x) = 0$$

$$\int_{-\infty}^{\infty} |y(x)|^2 dx < \infty$$

where p(x) is a sufficiently regular real valued function of the real variable x. In particular we are concerned with the asymptotic distribution of large positive eigenvalues for this problem which arises naturally in quantum mechanics. We note that the boundary condition, $\int_{-\infty}^{\infty} |y(x)|^2 dx < \infty$, implies that p(x) must have at least one zero. Recently, Weinberg [11], generalizing results of Sibuya [8] and Evgrafov and Fedoryuk [1], has obtained the asymptotic distribution of large positive eigenvalues for this problem under the assumptions that p is a real analytic function with an even finite number of simple real zeros that has polynomial type growth as $|x| \to \infty$

Clearly, we are again concerned with the asymptotic representations of solutions of a singularly perturbed differential equation on an interval with interior turning points. In essence this problem can be reduced to a connection problem. We p(x) is a polynomial, Evgrafov and Fedoryuk and Sibuya use the theory of irregular singular points at x = ∞ to study the properties of the Stokes' curves in the complex plane which enables them to solve the connection problem. The connection problem is solved for each turning point individually by Evgrafov and Fedoryuk while Sibuya and Weinberg solve the connection problem for pairs of turning points, greatly simplfying the resulting computations. In all cases, the problem can be reduced to studying the mapping

$$z = \int^{x} \sqrt{p(t)}\, dt$$

which is closely related to the Stokes curves.

By a confluence of simple turning points we may modify this technique so that it is applicable to the boundary value problem considered in

Section 2. However, generalizations to higher order turning point problems are not immediate.

5. Conclusion.

In this note we have discussed two physically motivated singular perturbation problems, a two point boundary value problem and an eigenvalue problem. Each problem represents the simplest such problem since the corresponding differential equation is of second order and the turning points are either simple or second order. However, these problems already tax our mathematical ability since the techniques for their solution represents current mathematical research. Further research in this direction is clearly needed.

BIBLIOGRAPHY

1. M.A. Evgrafov and M.V. Fedoryuk, "Asymptotic behavior as $\lambda \to \infty$ of the solution of the equation $w'' - p(z,\lambda)w = 0$ in the complex z - plane," Uspehi Mat. Nauk, 21 (1966), 3-50.

2. W.A. Harris, Jr, "Singular perturbations of two-point boundary problems," J. Math. Mech., 11 (1962), 371-382.

3. M. Iwano and Y. Sibuya, "Reduction of the order of a linear differential equation containing a small parameter," Kodai, Math. Seminar Reports, 15(1963), 1-28.

4. R.E. Langer, "On the construction of related differential equations," Trans. Am. Math. Soc., 81 (1956), 394-410.

5. R.Y. Lee, "On uniform simplification of Linear differential equations in a full neighborhood of a turning point," S. Math. Anal. Appl. 27 (1969), 501-510.

6. R.E. O'Malley, Jr., "On boundary value problems for a singularly perturbed differential equation with a turning point," SIAM J. on Appl. Math. (to appear).

7. Y. Sibuya, "Simplification of a system of linear ordinary differential equations about a singular point," Funkcial. Ekvac., 4 (1962), 29-56.

8. Y. Sibuya, "Subdominant solutions of the differential equation $y'' - \lambda^2 (x-a_1) \cdots (x-a_m) y = 0$," Acta Mat. to appear

9. W. Wasow, "On the asymptotic solution of boundary value problems for ordinary differential equations containing a parameter, J. Math. Phys., 32 (1944) 173-178.

10. W. Wasow, Asymptotic expansions for ordinary differential equations, John Wiley, New York, 1965.

11. L. Weinberg, Asymptotic distribution of eigenvalues for the boundary value problem $y'' - \lambda^2 p(x)y = 0$, $y \in L_2(-\infty,\infty)$, thesis, University of Minnesota, 1969.

STOCHASTIC CONTROL FOR SMALL NOISE INTENSITIES

Wendell H. Fleming
Brown University
Providence, Rhode Island

ABSTRACT

This paper is concerned with the approximate solution of some stochastic optimal control problems. The models of control systems which we consider are perturbations of the deterministic optimal control model of Pontryagin. We suppose that, in the stochastic problem, the system is perturbed by random disturbances which take the form of an additive white noise term in the system equations

$$d\xi = f(t,\xi,u)dt + \sigma dw,$$

where ξ,u denote the state and control processes. If the noise coefficient σ is small, it is reasonable to seek an approximate solution of the stochastic problem in terms of quantities computable from the Pontryagin problem (with $\sigma = 0$). Rather precise results of this kind are obtained for the case when the states $\xi(t)$ are completely observed by the controller at each time t.

Mathematically the problem becomes one of singular perturbation of the Hamilton-Jacobi equation for the Pontryagin problem (a partial differential equation of first order) by adding a small second order term. The objective is to expand the solution of the perturbed problem in powers of the noise variance $\varepsilon = \sigma^2$. Some results of this type are shown to hold in regions where the solution of the unperturbed (Hamilton-Jacobi) equation is sufficiently well behaved. The same method estimates how well the (closed loop) optimal control for the Pontryagin problem performs in the stochastic control problem. Examples show that the expansion may fail to be valid at points where the solution of the unperturbed problem misbehaves.

(This paper is available in preprint form as TR 70-1 Center for Dynamical Systems, Division of Applied Mathematics, Brown University.)

SINGULAR PERTURBATIONS AND MINIMUM FUEL SPACE TRAJECTORIES *

T.N. EDELBAUM
MIT Charles Stark Draper Laboratory
Cambridge, Massachusetts

ABSTRACT

A number of examples are given of the use of singular perturbation theory in space trajectory optimization. Singular perturbation problems have resulted from the use of four different small parameters. These parameters are 1) the mass of the launch, target, or swingby planet, 2) the reciprocal of the thrust for high thrust rockets, 3) the thrust for low thrust rockets, and 4) the reciprocal of the transfer time.

INTRODUCTION

The theory of minimum fuel space trajectories was originally developed by Lawden in the early nineteen fifties. For trajectories in three spatial dimensions the nonlinear equations of state are of seventh order.

$$\ddot{\bar{r}} = \bar{g}(\bar{r}, t) + \frac{\bar{f}}{m} \tag{1}$$

$$\dot{m} = \frac{f}{c} \tag{2}$$

$$0 \le f \le f_{max} \tag{3}$$

In these equations $\bar{r}$ is the position vector, $\bar{g}$ is the gravitational acceleration, $\bar{f}$ is the rocket thrust, m is the mass, and c is the constant exhaust velocity of the rocket. Lawden expressed the necessary conditions in terms of the adjoint vector for velocity, $\bar{\lambda}$, which he called the primer vector (Ref. 1). The Hamiltonian for this problem may be written as Eq. 4 where σ is the adjoint variable for mass.

$$H = \frac{f}{m}\left(\lambda - \frac{\sigma m}{c}\right) + \bar{\lambda} \cdot \bar{g} - \dot{\bar{\lambda}} \cdot \dot{\bar{r}} \tag{4}$$

In this equation the maximum principle has already been used to optimize the thrust direction by pointing it in the direction of the primer vector. The adjoint equations are given by equations (5) and (6).

$$\ddot{\bar{\lambda}} = \bar{\lambda} \cdot \frac{\partial \bar{g}}{\partial \bar{r}} \tag{5}$$

$$\dot{\sigma} = \frac{\lambda f}{m^2} \tag{6}$$

The Hamiltonian is linear in the thrust magnitude, so the optimal magnitude of the thrust depends on the sign of the switching function.

$$\lambda - \frac{\sigma m}{c} < 0 \qquad f = 0 \tag{7}$$

$$\lambda - \frac{\sigma m}{c} = 0 \qquad 0 \le f \le f_{max} \tag{8}$$

---

* The preparation of this paper was partially supported by NASA contract NAS-9-10386.

$$\lambda - \frac{\sigma m}{c} > 0 \qquad f = f_{max} \tag{9}$$

The optimal trajectory will usually consist of a sequence of coasting arcs and maximum thrust arcs. A typical time history of λ and $\sigma m/c$ for a fixed time trajectory with three thrusting periods is shown in the upper portion of Fig. 1. During the coasting arcs $\sigma m/c$ will be constant. The primer vector has a stationary maximum during the interior thrusting arc. This will always be true for an interior arc as the primer vector is continuous through several time derivatives (Eq. 5).

IMPULSIVE TRAJECTORIES

Lawden considered the case where f_{max} is allowed to become infinite, leading to impulsive changes in velocity. Across such a velocity impulse the mass will change by a finite amount given by Eq. 10.

$$\frac{m_2}{m_1} = e^{-\frac{\Delta V}{c}} \tag{10}$$

Lawden derived the necessary conditions for impulsive trajectories by a limiting argument (Ref. 1). It is interesting that he obtained the correct condition even though his work predated the maximum principle. In fact, the impulsive case is of such a nature that the standard maximum principle is not applicable. In recent years Lawden's results have been rigorously verified by Neustadt, Rishel, and Warga (Refs. 2, 3, 4).

If the thrust is allowed to become unbounded $\sigma m/c$ becomes a constant of the motion which may be taken as unity. The thrust is applied only when the primer vector has unit magnitude. The primer vector cannot exceed unity on an optimal trajectory. The Hamiltonian is only defined on the open interval from the initial time to the final time for fixed time problems. A typical primer vector history for a fixed-time three-impulse trajectory is shown in the lower part of Fig. 1.

Although the impulsive case can be treated by the extensions of the maximum principle in Refs. 2-4, Jack Warga has suggested a simple regularizing transformation which allows the ordinary form of the maximum principle to be used. Define a new independent variable τ and a new control variable u by Eqs. 11 and 12.

$$d\tau \equiv (1 + \frac{f}{mu^*})\,dt \tag{11}$$

$$u \equiv \frac{fu^*}{mu^* + f} \tag{12}$$

When u is equal to its maximum value u^* the thrust is infinite, Eq. 13.

$$\frac{f}{m} = \frac{u}{1 - \frac{u}{u^*}} \tag{13}$$

The Hamiltonian for the new variables is given by Eq. 14.

$$H_\tau = (\lambda - 1)u + (1 - \frac{u}{u^*})(\bar{\lambda} \cdot \bar{g} - \dot{\bar{\lambda}} \cdot \dot{\bar{r}}) \tag{14}$$

The conventional form of the maximum principle can be used for this new Hamiltonian. The Hamiltonian is now defined on the closed interval from the initial to the final time.

The extremals neighboring an optimum impulsive trajectory provide an interesting example of a singular perturbation problem. If the nominal extremal has two or more impulses, then neighboring extremals with the same number of impulses will exist (Ref. 5). However, near the end of any extremal only one impulse will remain on the nominal extremal. Neighboring extremals to this single impulse nominal trajectory may require as many as three small midcourse impulses to meet the desired terminal conditions. While the neighboring extremals only involve small variations in the state, they obviously require large variations in the adjoint equations. The primer vector magnitude will be required to go from a single endpoint maximum to as many as three initial and/or interior maxima and an endpoint maximum, all of unit magnitude.

If the nominal extremal has a finite bound on thrust magnitude and a single coasting arc followed by a single maximum thrust arc, then neighboring extremals with the same sequence of arcs will generally exist. However, the region covered by these neighboring extremals may be very small.

SINGULAR EXTREMALS

The case where the switching function remains identically zero over a finite time interval leads to what are known as singular arcs in both the classical calculus of variations and modern control theory. This usage of the term "singular" may be distinct from its usage in "singular perturbation theory". However, singular extremals do form a manifold of lower dimensionality than do the non-singular extremals so neighboring extremals do involve a singular perturbation.

The theory of singular extremals is developing rapidly at the present time (Refs. 6-9). Ref. 6 shows that junctions between singular and non-singular extremals for bounded thrust rocket trajectories are very complex. The junctions involve a well defined infinite sequence of maximum thrust and coasting arcs whose limit point is the beginning of the intermediate thrust singular arc. What is surprising is that if the bound on thrust is removed, then there can be a simple junction between an impulse and a singular arc. Robbin's demonstration of the existance of minimizing singular arcs in an inverse square field is of great theoretical interest and might eventually prove to have practical consequences (Ref. 10).

LARGE THRUST TRAJECTORIES

Impulsive thrust extremals are much easier to calculate than finite thrust extremals. As a result it is often desirable to calculate finite thrust trajectories as a singular perturbation of impulsive trajectories. Practical theories for this problem have been developed by Robbins, Hazelrigg and Lion, and Andrus, Refs. 11-13. These analyses are usable as long as the thrusting time is not too long. Their range of validity coincides with a large class of practical problems.

SWINGBY TRAJECTORIES

The masses of the planets are quite small compared to the mass of the sun. However, a close approach to a planet can strongly perturb a Heliocentric trajectory. These close planetary approaches can often decrease the flight time and/or the fuel consumption for interplanetary missions. A typical example is shown in Fig. 2. This is a 3.33 year trajectory to Saturn which passes Jupiter at 5 planetary radii after 1.55 years. Except for guidance corrections, the vehicle is in a free coasting orbit after leaving the Earth. A direct flight to Saturn would require a larger launch energy even if the flight time was greatly increased. The effect of Jupiter in this example is so large that it changes the heliocentric trajectory from an ellipse with an eccentricity of 0.786 to a hyperbola with an eccentricity of 2.10.

An accurate and useful theory has been developed for these problems by Breakwell and Perko (Ref. 14). This theory has been constructed by the method of matched asymptotic expansions. The inner solutions consist of planetcentric hyperbola's which are perturbed by the sun while the outer solutions consist of heliocentric conics perturbed by the planets. The solution contains terms of order $\epsilon^{1/2}$, ϵ, and $\epsilon^{3/2}$ and is in error by terms of order ϵ^{2}. The small parameter E is the ratio of the planetary mass to the solar mass.

This theory has been developed for the unpowered, ballistic case. However, it can easily be extended to handle optimum impulsive trajectories. The Breakwell-Perko theory can also handle lunar trajectories although the errors in this case are large enough to be significant.

A different singular perturbation approach to lunar trajectories was developed earlier by Kevorkian and various coworkers (e.g. Ref. 15). This theory uses a near rectilinear earth-moon trajectory as the outer solution. The accuracy of this theory is comparable to the Breakwell-Perko theory. It has recently been shown to yield many of the qualitative properties of earth-moon trajectories (Ref. 16).

LONG DURATION TRAJECTORIES

The theory of minimum fuel impulsive orbit transfer in an inverse square field has been extensively developed, particularly for the time open case (Ref. 17). For some cases the absolute minimum fuel solution requires infinite time while in other cases the time involved is finite. One case where the optimum transfer time is almost always infinite is transfer from an ellipse to a hyperbola (Ref. 18). In these cases a long transfer time may be used to reduce the fuel consumption close to the absolute minimum. A singular perturbation theory is being developed for these problems (Refs. 19, 20).

Figure 3 illustrates a four impulse transfer from an initial elliptic orbit to a specified hyperbolic asymptote. The first impulse transfers the vehicle from the initial ellipse onto a highly eccentric ellipse which approximates a straight line in the outer region. The second and third impulses are used to change the orientation of this highly eccentric ellipse and to reduce its perigee radius to the minimum allowable. A fourth impulse then transfers the vehicle onto the escape hyperbola.

The solution is developed as an asymptotic series in inverse third powers of the time between the first and fourth impulses. The magnitude of the first and fourth impulses involve terms of order unity and minus two-thirds. The magnitude of the second and third impulses involve terms of order minus one-third and minus two-thirds. The location and directions of the impulses are also developed as asymptotic series.

LOW THRUST TRAJECTORIES

One class of space propulsion devices is characterized by very small accelerations on the order of one ten-thousandth of the standard acceleration of gravity. For these electric propulsion devices, the thrust may be regarded as a small parameter, at least in the close vicinity of planets.

The method of averaging may be used to treat these problems (Refs. 21-23). A variation of parameters formulation is used with the conventional elliptic orbit elements being the slowly varying parameters. In the first approximation the order is reduced because the position in the orbit is averaged out. By going to the improved first approximation (Ref. 23) which considers first order periodic as well as secular terms, the orbital position can be reintroduced into the problem. Fig. 4 illustrates a minimum fuel, fixed time transfer between two coplanar elliptic orbits.

LOW THRUST ESCAPE TRAJECTORIES

If an electric propulsion system is to be used for interplanetary missions it might first be placed in a low altitude circular orbit and allowed to spiral slowly away from the earth or it might be injected onto a hyperbolic orbit which rapidly takes it into interplanetary space. Both possibilities are illustrated in Fig. 5.

The case of a circular initial orbit has been treated most thoroughly by Breakwell and Rauch (Refs. 24-26). They asymptotically match an analytical inner solution with a numerically determined intermediate solution and an analytical outer solution.

The case of a hyperbolic initial orbit has been treated by Melbourne and Sauer (Ref. 27), and by Edelbaum (Ref. 28). In this case the inner solution is an unpowered hyperbola. By approximating the trajectory as being essentially radial an analytic solution is obtained for the effect of the planet on the heliocentric trajectory.

REFERENCES

1. Lawden, D. F., Optimal Trajectories for Space Navigation, Butterworths, London, 1963.

2. Neustadt, L. W., "A General Theory of Minimum Fuel Space Trajectories", SIAM Journal on Control, Vol. 3, No. 2, 1965, pp. 317-356.

3. Rishel, R. W., "An Extended Pontryagin Principle for Control Systems whose Control Laws Contain Measures", SIAM Journal on Control, Vol. 3, No. 2, 1965.

4. Warga, J., "Variational Problems with Unbounded Controls", SIAM Journal on Control, Vol. 3, 1965, pp. 424-438.

5. Edelbaum, T. N., "Minimum Impulse Guidance", AIAA Journal, Vol. 8, No. 3, 1970, pp. 437-439.

6. Kelley, H. J., Kopp, R. E. and Moyer, H. G., "Singular Extremals", Chapter 3 of Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967.

7. McDanell, J. P. and W. F. Powers, "New Jacobi-Type Necessary and Sufficient Conditions for Singular Optimization Problems", AIAA Journal, Vol. 8, No. 8, 1970, pp. 1416-1420.

8. Jacobson, D. H., "Sufficient Conditions for Non-Negativity of the Second Variation in Singular and Nonsingular Control Problems",

9. Speyer, J. L., and Jacobson, D. H., "Necessary and Sufficient Conditions for Singular Control Problems, A Transformation Approach",

10. Robbins, H. M., "Optimal Rocket Trajectories with Subarcs of Intermediate Thrust", Proceedings XVII IAF Congress, Madrid 1966.

11. Robbins, H. M., "Analytical Study of the Impulsive Approximation", AIAA Journal, Vol. 4, No. 8, 1966, pp. 1417-1423.

12. Hazelrigg, G. A. and P. M. Lion, "Analytical Determination of the Adjoint Vector for Optimum Space Trajectories", AIAA Paper No. 69-916, August, 1969.

13. Andrus, J. F., "Linear and Higher Order Corrections to Optimum Multi-Impulse Trajectories", AIAA Paper No. 70-1014, August 1970.

14. Breakwell, J. and Perko, L., "Matched Asymptotic Expansions, Patched Conics and Interplanetary Trajectories", Progress in Astronautics and Aeronautics, Vol. 17, R. L. Duncombe, ed., Academic Press, New York 1966.

15. Lagerstrom, P. A. and Kevorkian, J., "Nonplanar Earth-to-Moon Trajectories in the Restricted Three-Body Problem", AIAA Journal, Vol. 4, No. 1, January 1966, pp. 149-152.

16. Lancaster, J. E., "Invariant Properties of Families of Moon-to-Earth Trajectories", to appear in Celestial Mechanics.

17. Gobetz. F. W. and Doll, J. R., "A Survey of Impulsive Trajectories", AIAA Journal, Vol. 7, No. 5, May 1969, pp. 801-834.

18. Edelbaum, T. N., "How Many Impulses?", Astronautics and Aeronautics, Nov. 1967, pp. 64-69.

19. Edelbaum, T. N., "Optimal Nonplanar Escape from Circular Orbits", AIAA Paper No. 70-1038, August 1970.

20. Edelbaum, T. N., "Optimal Nonplanar Escape from Elliptic Orbits", To be presented at AIAA Aerospace Sciences Meeting, January 1971.

21. Edelbaum, T. N.,"Optimum Thrust-Limited Orbit Transfer in Strong Gravity Fields", Lecture Notes in Mathematics 132, Symposium on Optimization, Balakrishman, et. al., ed., Springer-Verlag, Berlin, 1970.

22. Edelbaum, T. N., "Optimum Power-Limited Orbit Transfer in Strong Gravity Fields", AIAA Journal, Vol. 3, 1965, pp. 921-925.

23. Edelbaum, T. N., "An Asymptotic Solution for Optimum Power-Limited Orbit Transfer", AIAA Journal, Vol. 4, 1966, pp. 1491-1494.

24. Breakwell, J. V., and Rauch, H. E., "Optimum Guidance for a Low Thrust Interplanetary Vehicle", AIAA Journal, Vol. 4, 1966, pp. 693-704.

25. Breakwell, J. V., and Rauch, H. E., "Asymptotic Matching in Power-Limited Interpalnetary Transfers", AAS Science and Technology Series, Vol. 11, pp. 281-300.

26. Breakwell, J. V., and Rauch, H. E., "Three Dimensional Guidance for Low Thrust Escape from a Planet", Proceedings 19 IAF Congress, New York, 1968.

27. Melbourne, W. G., and Sauer, C. G., "Performance Computation with Pieced Solutions of Planetocentric and Heliocentric Trajectories for Low Thrust Missions", JPL Space Programs Summary, No. 37-36, Vol. IV, Dec. 1965.

28. Ragsac, R. V., et. al., "Study of Low Acceleration Transportation Systems", United Aircraft Research Laboratories Report D-910262-3, July 1965.

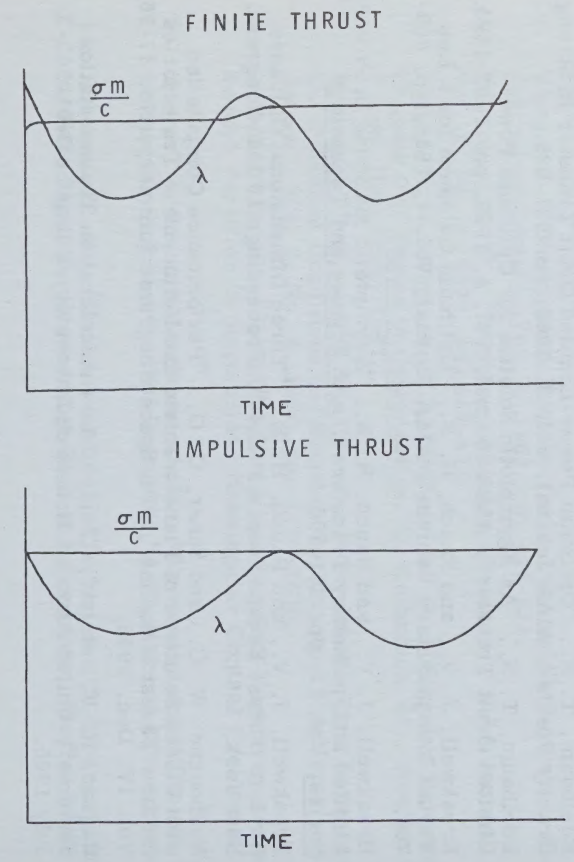


Figure 1

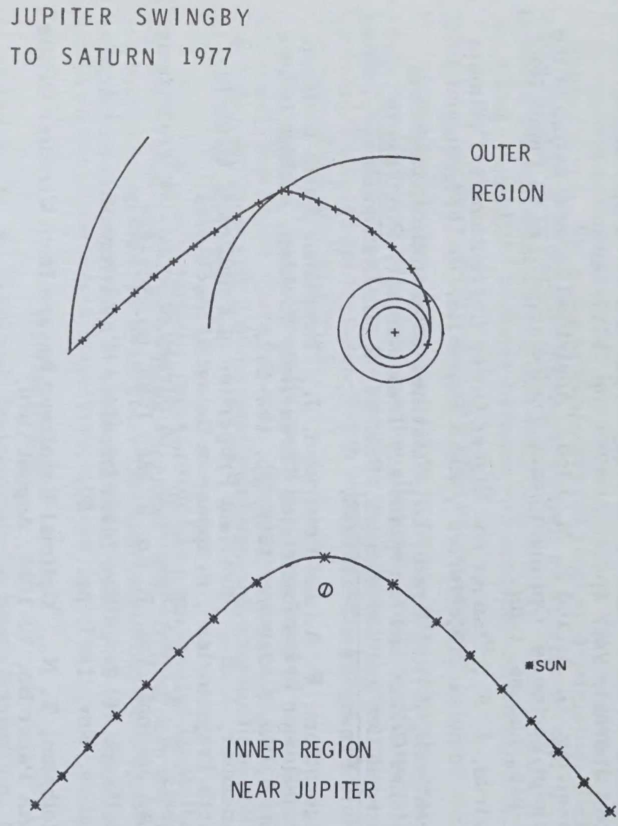

JUPITER SWINGBY
TO SATURN 1977
OUTER
REGION
SUN
INNER REGION
NEAR JUPITER

Figure 2

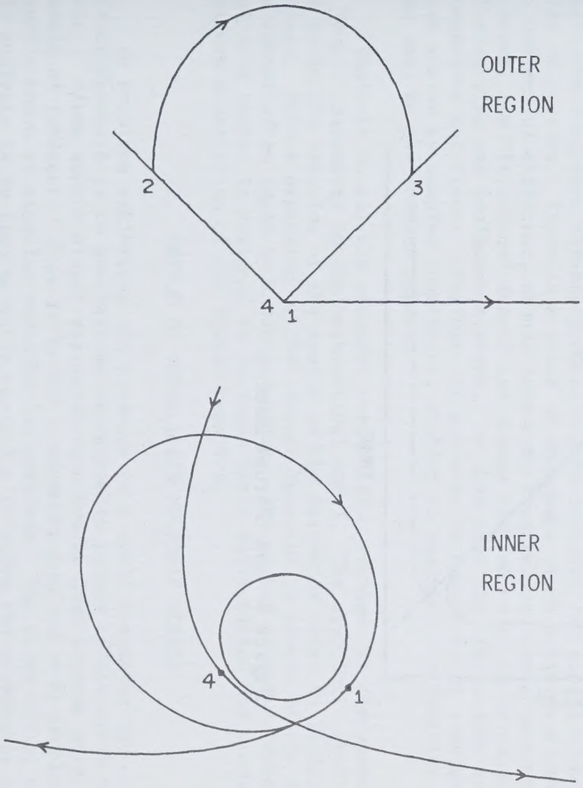

OUTER
REGION
2
3
4
1
INNER
REGION
4
1

Figure 3 Four Impulse Escape

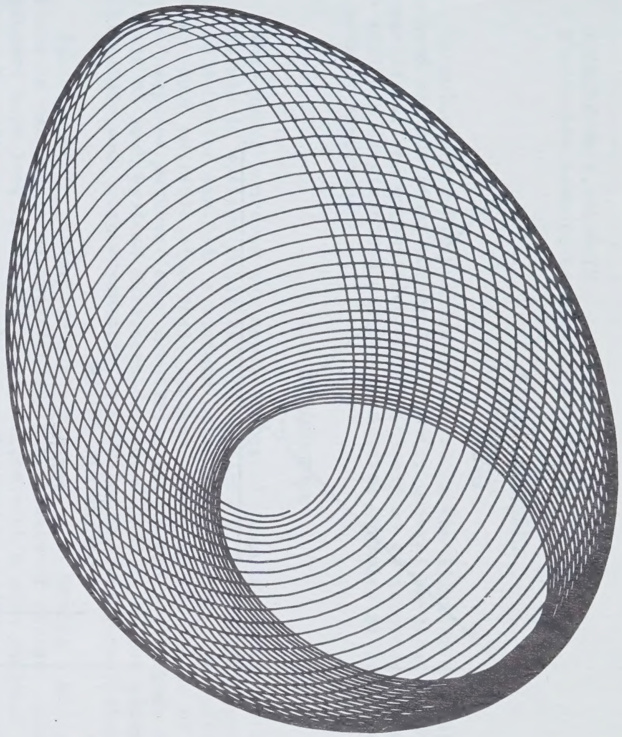


Figure 4 Low Thrust Transfer

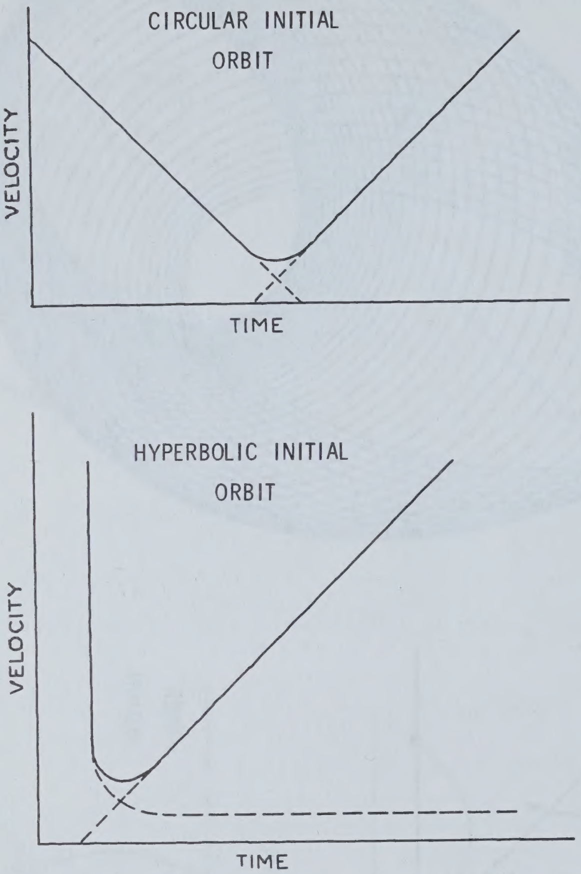


Figure 5 Low Thrust Escape

ANALYSIS OF SINGULAR PERTURBATIONS BY THE METHOD OF GENERALIZED MULTIPLE SCALES

RUDRAPATNA V. RAMNATH
Electronic Associates, Inc., Princeton, New Jersey

ABSTRACT

Singular perturbations are discussed in the light of multiple scales. The need for nonlinear scales for studying general physical phenomena is demonstrated. An extension of known principles of asymptotics facilitating the analysis is presented and the transition dynamics of V/STOL vehicles is briefly discussed.

1. INTRODUCTION

This paper concerns the applications of singular perturbation methods to problems in systems analysis. In particular, we consider systems whose dynamic behavior exhibits a mixture of motions occurring at different rates. Such multiple time scale systems abound in the fields of engineering and applied mathematics. However, rigorous and new methods for the analysis of such systems have been developed only in recent years. We shall mainly deal with asymptotic methods for analyzing such phenomena.

Examples of multiple time scale systems, i.e. having a mixture of motions at varying rates, exist in a great variety of fields. Chemical engineers have long known kinetic reactions with a wide range of reaction rates. Economic systems show rapid variations in addition to other changes which occur very slowly. The field of aerospace sciences is replete with such problems. A well-known example is the motion of an earth satellite. The orbital behavior shows a rapid motion similar to Keplerian motion, while the long term effects produce a slow rotation (or oscillation) of the line of apsides. There is a class of phenomena which exhibits a mixture of motions of different, though constant, frequencies. More generally, however, the "frequency" itself can change continuously, as is usual in time varying and nonlinear systems. In this case the time scales are themselves not linear. Further, it may even be necessary to consider the time scales as complex quantities, similar to analytic continuation from the real to the complex plane.

Implicit in all the singular perturbation problems is the presence of a "small" parameter in the mathematical model. The asymptotic problem is to study the behavior of the system as the parameter takes on limiting values. Direct expansion of the unknown function in powers of the small parameter often leads to nonuniformities. The method of generalized multiple scales is designed to overcome these difficulties and provide a uniform description of the system behavior.

2. METHOD OF GENERALIZED MULTIPLE SCALES

In problems exhibiting the presence of a small parameter the most direct approach is to generate an expansion in powers of the small parameter. Often such a direct representation breaks down in some part of the domain of interest. This is termed a nonuniformity and such problems are usually known as singular perturbation problems. The major aim of asymptotic analysis is to find suitable variables in which the problem is amenable to a direct perturbation approach, i.e. has a meaningful limiting case.

Methods have been devised to overcome the difficulty of nonuniformities and render the approximations uniformly valid. In this connection we may note the important work of Poincaré [1] in expanding the amplitude and frequency of nonlinear oscillatory systems of celestial mechanics, Lighthill's [2] expansion of the independent variable (PLK method), the method of matched asymptotic expansions of Lagerstrom, Kaplun [4], Kruskal [5] and others. Another important method that has been developed recently is the method of multiple scales.

The method is particularly useful when a physical phenomenon exhibits qualitatively different behaviors when observed in a fine or a gross sense. The method enables us to extract the different behaviors individually. The origin of the multiple scales method can be traced to the works of Poincaré on secular expansions in celestial mechanics and to the works of Krylov, Bogoliubov and Mitropolsky [6], who allowed the constants appearing in direct perturbation theory to be slowly varying functions. Cole and Kevorkian [7] have studied it in connection with some nonlinear systems and Frieman and Sandri in the context of statistical mechanics. Cochran and Nayfeh [4] have also applied this idea to a number of physical problems. Most of the early applications considered linear time scales, i.e. the new scales are linear functions of the original independent variable. For a large class of problems, however, linear scales are inadequate and a more general concept is necessary. The theory was generalized by Ramnath [8,9] and Sandri to include nonlinear scales, which could, in general, be complex quantities.

The method itself consists of reparameterizing the system equations by extending the domain of the independent variable into a space of higher dimension. The range of the dependent variable could also be extended, if necessary. The ordinary differential equations describing the system dynamics are converted into partial differential equations. The parameter is used to obtain a hierarchy of equations ordered according to their importance. These partial differential equations are solved order by order such that their solutions coincide with the solutions of the original equations along certain lines in the extended space, known as "trajectories." We have two degrees of freedom: the choice of the trajectories and the choice of the extension itself. This freedom is utilized in obtaining a simpler and a smoother dependence of the solution upon the parameter.

We may now consider a definition of extension: Given a function $f(t)$ where t is, in general, an n dimensional vector, and a function $\underline{f}(\tau_1, \tau_2, \dots \tau_N)$ of the N independent variables $\tau_1, \dots \tau_N$ (each of which is an n dimensional vector), $\underline{f}$ is said to be an extension of f if and only if there exists a set of $N \times n$ equations $\tau_n = \tau_n(t)$, $n = 1, 2 \dots N$ which, when inserted into $\underline{f}$ give: $\underline{f}(\tau_1(t), \tau_2(t), \dots \tau_N(t)) = f(t)$. The space of the N-tuplets $\underline{\tau} = \{\tau_1, \tau_2, \dots \tau_N\}$ is called the extension of the domain and $\tau_n(t)$ are called the "trajectories" in the extended domain.

3. APPLICATION TO DIFFERENTIAL EQUATIONS

We will now apply these ideas to ordinary differential equations. The basic work on this study is given in [8,9]. We will first show the inadequacy of linear scales even in the simple first order linear equation.

3.1 First Order Equations

Consider the equation:

$$y' + \varepsilon\,\omega(t)\,y = 0 \qquad (3.1.1)$$

A straight perturbation expansion is nonuniform. Let us therefore extend:

$$\left.\begin{aligned} t &\to \{\tau_o, \tau_1\}\ ;\ \tau_o = t,\ \tau_1 = \varepsilon t \\ y(t, \varepsilon) &\to \bar{y}(\tau_o, \tau_1) \end{aligned}\right. \qquad (3.1.2)$$

We denote differentiation w.r.t τ_o and τ_1 by the subscripts 1 and 2 respectively e.g. $\frac{\partial \bar{y}}{\partial \tau_o} \equiv \bar{y}_1$. The extended perturbation equations are:

$$\left.\begin{aligned} \bar{y}_1 &= 0 \qquad (a) \\ \bar{y}_2 + \omega(\tau_o)\bar{y} &= 0 \qquad (b) \end{aligned}\right\} \qquad (3.1.3)$$

Integration of (3.1.3 a) gives $\bar{y}(\tau_o, \tau_1) = A(\tau_1)$ (3.1.4)

(3.1.3 b) yields $\bar{y}(\tau_o, \tau_1) = c \exp(-\omega\,\tau_1)$ (3.1.5)

We "restrict" the extended solution (3.1.5) along the trajectories as:

$$\tilde{y}(t, \varepsilon) = c \exp(-\varepsilon\omega t) \qquad (3.1.6)$$

which is not a solution of (3.1.1) unless ω is a constant.

Let us instead consider the extension:

$$t \to \{\tau_o, \tau_1\};\ \tau_o = t,\ \tau_1 = \varepsilon k(t) \qquad (3.1.7)$$

where $k(t)$ is a "clock" function, as yet undetermined. The extended equations now are:

$$\left.\begin{aligned} \bar{y}_1 &= 0 \qquad (a) \\ \dot{k}\,\bar{y}_2 + \omega\,\bar{y} &= 0 \qquad (b) \end{aligned}\right\} \qquad (3.1.8)$$

Thus, $\bar{y}(\tau_o, \tau_1) = A(\tau_1)$ and using this in (3.1.8 b)

$$-A'/A = \omega/\dot{k} = \text{constant} = 1 \qquad (3.1.9)$$

Therefore, from these equations we obtain:

$$\tilde{y}(t, \varepsilon) = c \exp\left(-\varepsilon \int^{t} \omega(s)\,ds\right) \qquad (3.1.10)$$

which is an exact solution of (3.1.1).

We note that the clock $k(t)$ is a highly nonlinear function depending on $\omega(t)$. The function surface in the extended space is depicted in Fig. 1. The restriction of the surface along trajectory $\tau_1 = \epsilon k(\tau_o)$ describes the solution in t.

A similar approach is taken in the study of higher order equations also. However, now we seek asymptotic solutions as the exact solutions cannot be found in general.

3.2 Second Order Equations

Consider the equation:

$$y'' + \omega_1(\epsilon t)y' + \omega_o(\epsilon t)y = 0 \tag{3.2.1}$$

where $0 < \epsilon << 1$. This equation with slowly varying coefficients can be rewritten by choosing a new independent variable $\tau = \epsilon t$ as:

$$\epsilon^2 y'' + \epsilon\omega_1(\tau)y' + \omega_o(\tau)y = 0 \tag{3.2.2}$$

The task of finding asymptotic solutions in the limit $\epsilon \to 0$ constitutes a singular perturbation problem. This can also be written as:

$$y'' + \lambda\omega_1 y' + \lambda^2\omega_o y = 0 \tag{3.2.3}$$

where $\lambda = \frac{1}{\epsilon} >> 1$. Finally, as the first derivative term can always be transformed out it is convenient to write the canonical form:

$$y'' + \epsilon\,\omega(t)y = 0 \tag{3.2.4}$$

Solutions are sought in the limits of small and large ϵ.

An important feature of asymptotic analysis, in general, is the reduction of a complex system into a solvable subsystem. The analysis of the subsystem, which is usually simpler than the original system, thus facilitates the analysis of the complete system. If done properly, the subsystem solution represents the true solution in some limit and this property usually leads to significant simplifications in systems analysis. This is the Principle of Simplification: An asymptotic (or asymptotological analysis tends to simplify the system considered [4].

In studying this problem we shall invoke the principle of simplification. This is done basically by neglecting terms containing the small parameter. But usually the relative asymptotic magnitude of the terms to be compared depends upon some knowledge not yet available or on some assumption or decision not yet made. For maximal flexibility and generality we should keep all the terms, i.e. consider that they could all be of comparable magnitude. Any term that is definitely known to be smaller than others can be neglected but no term should be neglected unless it is indeed small. Pending a decision an attempt to balance the competing terms leads to the correct ordering. The subsystem that contains the most information is the one containing the maximum number of comparable terms; it simplifies the least. This is the Principle of Minimal Simplification (or Maximal Balance). If there is more than one maximal subsystem and they are independent of each other and noninclusive (as is often the case), each maximal subsystem corresponds to a different asymptotic behavior.

These ideas were first explicitly expressed by Kruskal [4] although they were familiar to those working in asymptotics. He noted that in dealing with algebraic equations and conventional perturbation theory of differential equations the maximal sets can be determined by trying out different transformations of the form $x = \epsilon^m \xi$ and choosing different values for m. In each case higher order terms are neglected. The values of m for which the maximum number of terms are retained, describe different asymptotic behaviors. Alternatively, Kruskal suggests a simple graphical technique to determine the maximal ordering. We shall present an extension of this idea in the multiple scales theory [7].

In deriving solutions to (3.2.2) the variables are extended as:

$$t \rightarrow \{\tau_o, \tau_1\}; \quad \tau_o = t; \quad \tau_1 = \epsilon^m k(t) \tag{3.2.5}$$

m is chosen such that the most informative subsystem is obtained. The extended equation is:

$$\epsilon^2(y_{11}) + \epsilon^{2+m}\left(\ddot{k}y_2 + 2\,\dot{k}y_{12}\right) + \epsilon^{2+2m}\left(\dot{k}^2 y_{22}\right) + \epsilon(\omega_1 y_1) + \epsilon^{1+m}\left(\omega_1 \dot{k} y_2\right) + (\omega_o y) = 0 \tag{3.2.6}$$

The terms are all of the form ϵ^{a+bm}. In order to determine the maximal subsystems, the terms are represented as graphed points (Fig. 2) with a along the ordinate and b along the abscissa. Our extension [7] of the Principle of Minimal Simplification states that for $\epsilon << 1$, the most informative ordering is given by the lower convex support line of the set of graphed points. That is, a line is kinesthetically brought from below such that it passes through a maximum number of graphed points with no point lying below the line. The desire to balance the collinear points leads to a choice of m, which, in this case, is given by $m = -1$. The points in the shaded region are of higher order. Similarly for $\epsilon >> 1$, the upper convex support line determines the maximal balance. Therefore, the hierarchy of equations is given by:

$$\begin{aligned} \dot{k}^2 y_{22} + \omega_1 \dot{k} y_2 + \omega_o y &= 0 \qquad \text{(a)} \\ \ddot{k} y_2 + 2\dot{k} y_{12} + \omega_1 y_1 &= 0 \qquad \text{(b)} \\ y_{11} &= 0 \qquad \text{(c)} \end{aligned} \tag{3.2.7}$$

(3.2.7 a, b) can be solved [8] in the form:

$$\begin{aligned} y(\tau_o, \tau_1) &= \alpha(\tau_o)\,\beta(\tau_1) \\ &= \left(2\dot{k} + \omega_1\right)^{-\frac{1}{2}} \exp\left(\int \frac{\dot{\omega}_1 d\tau_o}{2(2\dot{k} + \omega_1)}\right) \exp(\tau_1) \end{aligned} \tag{3.2.8}$$

where the clock function k satisfies the equation

$$k^2 + 2\omega_1 k + \omega_o = 0 \tag{3.2.9}$$

If the equation is studied in the form (3.2.4) then the upper convex support line determines the maximal ordering. A similar approach can be followed even for linear equations of the nth order with slowly varying coefficients [8]. The Principle of Minimal Simplification leads to the proper ordering and the hierarchy of perturbation equations.

3.3 Boundary Layer Model

A model of a matching type of singular perturbation is given by:

$$\varepsilon y'' + Ay' + By = 0 \tag{3.3.1}$$

where $0 < \varepsilon << 1$; $y(o) = 0$, $y'(o) = c$. The solution has a rapidly varying part (inner solution) and a slowly varying part (outer solution). Conventionally, these are obtained separately and are then matched in the region of overlap or common boundary to yield a uniform solution. Some degree of art is involved in this process. However, using the multiple scales theory, the derivation of the inner and outer expansions and the need to match them are avoided. They are, instead, obtained as successive terms of one expansion. For example, the extension $\tau_{-1} = \frac{t}{\varepsilon}$; $\tau_i = \varepsilon^i t$, $i = 0,\ldots n$ yields the inner solution in the first order. Corrections are generated by considering higher order terms in the same expansion. The analysis is rendered especially clear by the Principle of Minimal Simplification.

4. APPLICATION TO VTOL DYNAMICS

The analysis of the dynamic motion of VTOL aircraft during the transition from hover to forward flight is rendered straightforward when viewed in the light of the multiple scales theory. The motion of the vehicle on the fast and slow scales can be separated systematically [7,9].

The VTOL vehicle necessarily has some means of tilting the thrust during the hover-forward flight transition. A tilt-wing vehicle is considered in this analysis. However, the method is applicable to other types of VTOL vehicles also. At hover the vehicle behaves like a helicopter and is dynamically unstable. The vehicle then accelerates until it reaches cruising velocity at which time it is dynamically stable. During transition the flight condition changes continuously. The vehicle dynamics change from instability to stability or vice versa. The vehicle is assumed to be trimmed throughout the transition. That is, the time histories of the forces and moments needed to trim the vehicle are assumed to be known so that they could be programmed in advance. The equations of motion are now linearized about a particular transition trajectory. The vehicle can be considered to be passing through a series of equilibrium states. Only motion in the plane of symmetry is considered. Further, the natural dynamics of the vehicle about a prescribed trajectory are studied.

The study of conventional aircraft at steady flight has been very successful in leading to an understanding of the vehicle motion and, further, to identify parameters important to desirable handling qualities.

For a VTOL vehicle a similar study has not been possible easily, mainly because of the difficulty of developing analytical solutions to the time varying equations. The multiple scales theory helps to solve this problem in a systematic manner. In this paper we will present only some results. The problem is examined in greater detail in [7,9].

The linearized homogeneous equations of motion for the longitudinal mode are written as:

$$\left.\begin{aligned} u' - X_u u - X_w w + g\,\theta &= 0 \qquad \text{(a)}\\ w' - Z_w w - Z_u u - V\,\theta' &= 0 \qquad \text{(b)}\\ \theta'' - M_q \theta' - M_u u - M_w w &= 0 \qquad \text{(c)} \end{aligned}\right\} \qquad (4.1)$$

The stability derivatives X_u, X_w... are predominantly functions of the flight velocity V though in general they depend on the wing tilt angle, thrust setting, etc. For a particular variation of the velocity with time (which is at the control of the pilot) the coefficients of the system equations become time dependent.

The "frozen" approximation, which considers the coefficients to be essentially constants, fails to describe the motion satisfactorily as it breaks down even within a cycle of the oscillation. The multiple scales theory separates the fast and slow motions as follows.

First the two degree of freedom case, valid near hover, is considered. The equations are:

$$\left.\begin{aligned} u' - X_u u + g\,\theta &= 0 \qquad \text{(a)}\\ w' - Z_w w &= 0 \qquad \text{(b)}\\ \theta'' - M_q \theta' - M_u u &= 0 \qquad \text{(c)} \end{aligned}\right\} \qquad (4.2)$$

We substitute for the stability derivatives as follows: $-X_u = 02$; $-Z_w = 0.1 + 0.004V$; $-Z_u = 0.25V/(10 + V)$; $-M_q = 0.1 + 0.0034V$; $-M_u = 0.015(-1 + V/150)$ and (i) $-M_w = (-0.02 + 0.00025V)V/150$; (ii) $M_w = 0.005 + 0.015(V/150)^2$. On decoupling the u and θ equations, we can write:

$$(1 + 0.1t)u^{(3)} + (0.3 + 0.081t)u^{(2)} + (0.02 + 0.0122t)u' + 0.48u = 0 \qquad (4.3)$$

The rapid variation in the u motion is given by $\exp(\tau_1)$ where the clock satisfies the "characteristic" equation:

$$(1 + 0.1t)\dot{k}^3 + (0.3 + 0.081t)\dot{k}^2 + (0.02 + 0.0122t)\dot{k} + 0.48 = 0 \qquad (4.4)$$

The roots of this equation are plotted and the solution is given in the form $\alpha(\tau_o) \exp(\tau_1)$ (Fig. 3), which represents the motion uniformly.

For the full three degree-of-freedom case the equations are much more complicated. The multiple scales theory still provides a uniform description of the motion. Typical root configuration and response are shown in Fig. 4. A detailed derivation of the solutions is given in Refs. 7, 9.

It is seen that two of the characteristic roots coalesce during the transition, leading to a turning point problem. The nature of the solution changes from monotonic to oscillatory behavior (for that mode) or vice versa. The asymptotic approximation, which is in terms of elementary functions, fails near the turning point. Continuation of the asymptotic solution derived on one side to the other side of the turning point will result in an error in phase. For the study of the natural dynamics of VTOL vehicles, precise phase information is not necessary as it merely represents the aircraft orientation in space. Stability and frequency of the motion are much more important and these are predicted by the multiple scales solution. If indeed a continuous solution is needed throughout the transition, specific connection formulae can be developed by studying the Stokes' phenomenon at the turning point. Such a connected solution could be then used for feedback control of the time varying VTOL vehicle. For greater mathematical elegance, one may resort to a uniform representation using higher transcendental functions in the manner of Langer, which would be valid throughout the transition.

5. CONCLUDING REMARKS

The method is seen to be mathematically rigorous and physically meaningful. The concept of nonlinear scales has been successfully applied to a number of nonlinear problems. Computationally the method seems to be very promising in separating the different motions and thus developing efficient computation and simulation schemes. In studying boundary layer phenomena by the multiple scales method, the need for matching different solutions is completely avoided. The "inner" and outer solutions are generated systematically by one expansion, order by order. In systems analysis, therefore, the method seems to be widely applicable and it is hoped that wider exposure of the technique will lead to significant advances in systems theory.

REFERENCES

1. Poincaré, Les Méthodes Nouvelles de la Mécanique Céleste.
2. Lighthill, M. J., "A Technique for Rendering Approximate Solutions to Physical Problems Uniformly Valid," Phil. Mag., 1949.
3. Van Dyke, M., Perturbation Methods in Fluid Mechanics, Academic Press.
4. Kruskal, M. D., "Asymptotology," MATT-160, Princeton University, 1962.
5. Bogoliubov and Mitropolsky, Asymptotic Methods in the Theory of Nonlinear Oscillations, Gordon and Breach, N. Y.
6. Cole, J. D., Perturbation Methods in Applied Mathematics, Blaisdell, 1968.
7. Ramnath, R. V., "A Multiple Time Scales Approach to the Analysis of Linear Systems," Ph.D. Dissertation, Princeton University, Princeton, N. J., 1967.
8. Ramnath, R. V. and Sandri, G., "A Generalized Multiple Scales Approach to a Class of Linear Differential Equations," J. Math. Anal.&Appl. Vol. 28, No. 2, Nov., 1969.

9. Ramnath, R. V., "Transition Dynamics of VTOL Aircraft," AIAA J., Vol. 8, No. 7, July, 1970.
10. Ramnath, R. V., "A New Analytical Approximation to the Thomas-Fermi Model in Atomic Physics," J. Math. Anal.&Appl., Vol. 31, No. 2, Aug., 1970.
11. Ramnath, R. V., "On a Class of Nonlinear Differential Equations of Astrophysics," J. Math Anal.&Appl. (to appear).

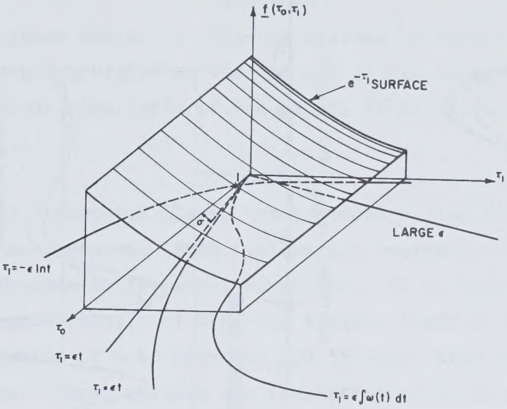

$f(\tau_0,\tau_1)$
$e^{-\tau_1}$ SURFACE
τ_1
LARGE ϵ
$\tau_1=-\epsilon \ln t$
σ
τ_0
$\tau_1=\epsilon t$
$\tau_1=\epsilon t$
$\tau_1=\epsilon\int\omega(t)\,dt$

FIGURE I. FUNCTION SURFACE IN EXTENDED SPACE

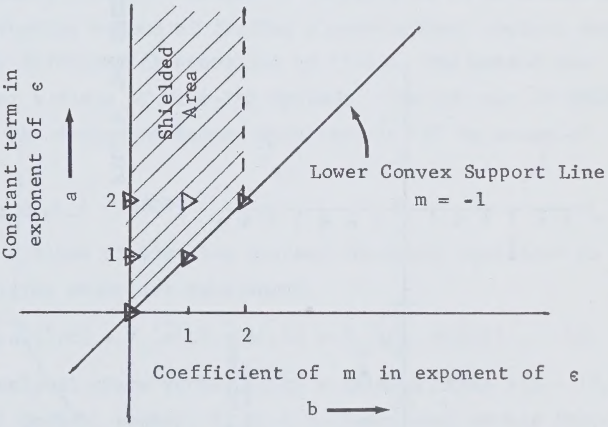


FIG. 2 BALANCING TERMS

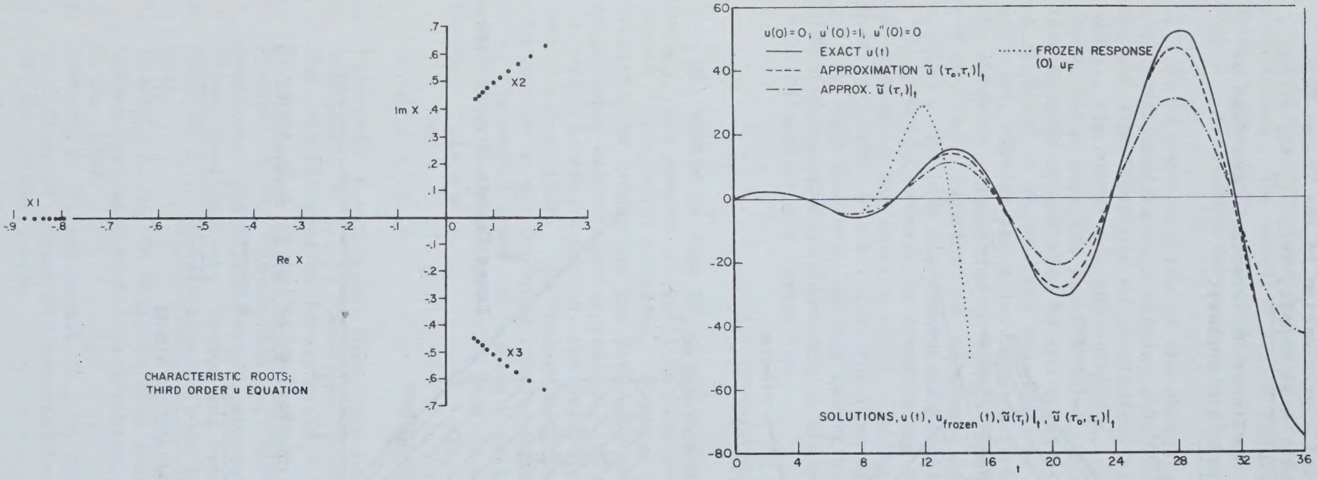


FIG. 3 TWO DEGREE OF FREEDOM CASE

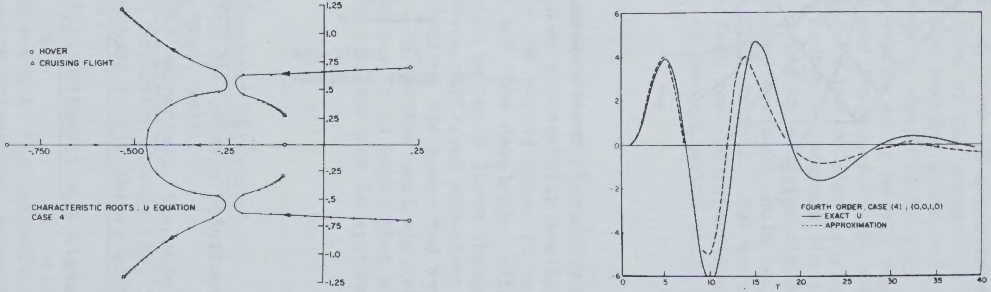


FIG. 4 THREE DEGREE OF FREEDOM CASE

SINGULAR PERTURBATION DESIGN OF A CLASS OF TIME-LAG SYSTEMS

P. SANNUTI and P. B. REDDY
Department of Electrical Engineering
Rutgers University
New Brunswick, New Jersey

ABSTRACT

A near-optimum design of time-lag systems by singular perturbation method has been formulated earlier by one of the authors. The computational and analytical simplicity of the method is illustrated here on a nonlinear example.

INTRODUCTION

In several industrial plants transportation lags or pure time delays are commonly encountered. Such systems are represented mathematically by a set of difference-differential equations. It is well known that the optimal control synthesis of time-lag systems leads to a boundary value problem consisting of both advanced and retarded type difference-differential equations. These systems are characterized to be equivalent to infinite order equations. As the computational difficulties of finding an optimal control increase with system order, a method of finding a near optimal control without requiring prohibitive amount of computation is of practical use. A singular perturbation method of finding a near-optimal control was formulated for ordinary differential equations in [1-4]. The method has been extended in [5] for a class of time-lag systems. The purpose of this paper is to illustrate the design procedure developed in [5] by means of an example.

THEORY

In this paper only a class of time-lag systems which are nonlinear in state but linear in delayed state are considered,

$$\dot{\tilde{x}}(t) \equiv f(\tilde{x}(t),\tilde{y}(t),\tilde{u}(t),t) = f_1(\tilde{x}(t),\tilde{u}(t),t) + f_2(\tilde{x}(t),t)\tilde{y}(t) \qquad (1)$$

where $\tilde{x}(t)$ is a n-dimensional state vector, $\tilde{y}(t)$ = delayed state $\tilde{x}(t - \tau)$, $\tilde{u}(t)$ is a r-dimensional control vector, f_1 is a n-dimensional vector function, f_2 is a n × n dimensional matrix function and τ is a small and positive time delay.

It is desired to find a control $\tilde{u}^*(t,\tau)$ satisfying the constraints (1) and minimizing the performance index,

$$J = \int_{t_0}^{T} V(\tilde{x}(t),\tilde{u}(t),t)dt, \qquad (2)$$

with the initial state,

$$\tilde{x}(t) = \phi(t);\ t_0 - \tau \leq t \leq t_0 , \tag{3}$$

and the final state $\tilde{x}(T)$ free.

The above optimization problem generates a boundary value problem consisting of both advanced and retarded type of difference-differential equations. Numerical solution of these equations becomes impractical if the order of the system (1) is high. Also a small time-delay represents a fast mode of the system. The standard methods of step-wise integration require the use of step sizes which are far below the response time of the fastest mode of a system. Because of this a direct design of systems having fast modes becomes economically impractical. A near-optimum design method has been proposed in [5], which deals with small time delays indirectly. It is shown there that the optimal control $\tilde{u}^*(t,\tau)$ can be expanded in MacLaurin series in τ. The first two terms of the series define the near-optimal control $\hat{u}^*(t,\tau)$,

$$\hat{u}^*(t,\tau) = \tilde{u}^*(t,0) + \left.\frac{\partial\tilde{u}(t,\tau)}{\partial\tau}\right|_{\tau = 0} \tau . \tag{4}$$

Thus the near-optimal control is obtained in two steps.

Step 1. The first term of the series, $\tilde{u}^*(t,0)$, corresponds to the control obtained when the time-lag τ is neglected and is termed as nominal control. This is obtained as a solution of the following optimization problem. Minimize the performance index,

$$J = \int_{t_0}^{T} V(\underline{x},\underline{u},t)dt , \tag{5}$$

subject to the constraints,

$$\dot{\underline{x}}(t) = f_1(\underline{x}(t), \underline{u}(t), t) + f_2(\underline{x}(t),t)\underline{x}(t) \tag{6}$$

with the initial state,

$$\underline{x}(t_0) = \phi(t_0) \tag{7}$$

and $\underline{x}(T)$ is free.

The under bar (_) implies that the variables correspond to the design when time-lag $\tau = 0$. In what follows, this design is called the nominal design.

The solution of this problem is well discussed in literature, say [6,7]. As can be noted only a part of the initial state, namely $\phi(t_0)$, is used here.

Step 2. In this step a procedure to find the control sensitivity $\frac{\partial\tilde{u}(t,\tau)}{\partial\tau}$ is outlined. Define the optimal sensitivity functions,

$$\tilde{\omega} = \frac{\partial\tilde{x}(t,\tau)}{\partial\tau}\Big|_{\tau = 0}, \quad \tilde{\upsilon} = \frac{\partial\tilde{u}(t,\tau)}{\partial\tau}\Big|_{\tau = 0} \tag{8}$$

As shown in [5], $\tilde{\omega}$ and $\tilde{\upsilon}$ are the solution of the equations,

$$\tilde{\upsilon} = -H^{-1}_{\underline{u}\,\underline{u}}\left((H_{\underline{x}\,\underline{u}} - f'_{\underline{u}}M)\,\tilde{\omega} + f'_{\underline{u}}\,\gamma\right) \tag{9}$$

$$\dot{\tilde{\omega}} = (A_1 - A_2M)\,\tilde{\omega} + A_2\,\gamma + \tilde{\delta}_1 \tag{10}$$

with the initial condition,

$$\tilde{\omega}(t_0) = \frac{1}{\tau}\, f_2(\phi(t_0),t_0)\int_{t_0}^{\tau}(\phi(t_0 - \alpha) - \phi(t_0))d\,\alpha \tag{11}$$

where,

M is the solution of a n × n dimensional symmetric matrix Riccati equation,

$$\dot{M} = -MA_1 - A_1'M + MA_2M - A_3\ ;\ M(T) = 0 \tag{12}$$

γ is the solution of a n-dimensional linear vector equation,

$$\dot{\gamma} = (MA_2 - A_1')\gamma + M\tilde{\delta}_1 + \tilde{\delta}_2\ ;\ \gamma(T) = 0 \tag{13}$$

H is the Hamiltonian function $= -V(\tilde{x},\tilde{u},t) + p'(f(\tilde{x},\tilde{y},\tilde{u},t))$ (14)

$$A_1 = f_{\underline{x}} + \underline{f}_2 - f_{\underline{u}}H^{-1}_{\underline{u}\,\underline{u}}H_{\underline{x}\,\underline{u}}$$

$$A_2 = -f_{\underline{u}}H^{-1}_{\underline{u}\,\underline{u}}f'_{\underline{u}}$$

$$A_3 = -H_{\underline{x}\,\underline{x}} - H_{\underline{x}\,\underline{y}} - H'_{\underline{x}\,\underline{y}} + H_{\underline{u}\,\underline{x}}H^{-1}_{\underline{u}\,\underline{u}}H'_{\underline{u}\,\underline{x}}$$

$$\tilde{\delta}_1 = -\underline{f}_2\dot{\underline{x}}$$

$$\tilde{\delta}_2 = H'_{\underline{x}\,\underline{y}}\dot{\underline{x}} - (\dot{\underline{f}}'_2 p)$$

Under bar (_) implies that the variables of Step 1 are used in evaluating the partials $f_x = \frac{\partial f}{\partial x}$, etc., p is the costate variable corresponding to x.

It should be stressed that this step takes into account the initial state $\phi(t)$, $t_0 - \tau \le t < t_0$, in order to obtain the initial condition for the state sensitivity $\tilde{\omega}(t)$.

EXAMPLE

Let,

$$\dot{\tilde{x}}_1 = (1 - \tilde{x}_1^2 - \tilde{x}_2^2)\, \tilde{x}_1(t - \tau) - \tilde{x}_2(t - \tau) + \tilde{u}\ ,$$

$$\dot{\tilde{x}}_2 = \tilde{x}_1(t - \tau)\ ,$$

$$J = \int_0^5 (\tilde{x}_1^2 + \tilde{x}_2^2 + 0.2\ \tilde{u}^2)\ dt\ ,$$

and the initial conditions be,

$$\tilde{x}_1(t) = 10\ t\ ; \qquad - \tau \leq t \leq 0$$

$$\tilde{x}_2(t) = 2 - 10\ t\ ; \quad - \tau \leq t \leq 0\ .$$

In the first step of the method, the nominal control $\tilde{u}^*(t,0)$ is obtained by solving the set of equations corresponding to nominal design, (5), (6) and (7). Then to obtain the control sensitivity $\tilde{\upsilon}$ the equations (9), (10), (12) and (13) are solved with the condition (11).

$$A_1 = \begin{bmatrix} (1 - 3\ \underline{x}_1^2 - \underline{x}_2^2) & -(1 + 2\ \underline{x}_1\underline{x}_2) \\ 1 & 0 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 2.5 & 0 \\ 0 & 0 \end{bmatrix}$$

$$A_3 = 2 \begin{bmatrix} 1 + 3\ \underline{p}_1\underline{x}_1 & \underline{p}_1\underline{x}_2 \\ \underline{p}_1\underline{x}_2 & 1 + \underline{p}_1\underline{x}_1 \end{bmatrix}$$

$$\tilde{\delta}_1 = \begin{bmatrix} -(1 - \underline{x}_1^2 - \underline{x}_2^2)\ \dot{\underline{x}}_1 + \dot{\underline{x}}_2 \\ - \dot{\underline{x}}_1 \end{bmatrix}$$

$$\tilde{\delta}_2 = \begin{bmatrix} \dot{\underline{p}}_1(1 - \underline{x}_1^2 - \underline{x}_2^2) + \dot{\underline{p}}_2 - 2\ \underline{x}_2\ \underline{p}_1\ \dot{\underline{x}}_2 \\ 2\ \underline{x}_2\ \underline{p}_1\ \dot{\underline{x}}_1 - \dot{\underline{p}}_1 \end{bmatrix}$$

where $\underline{x}_1$, $\underline{x}_2$, $\underline{p}_1$, $\underline{p}_2$ are the variables obtained in the nominal design.

The optimal control is obtained now so that the near-optimum and optimum designs can be compared. The optimal control is the solution of the following boundary value problem.

The state equations are,

$$\dot{\tilde{x}}_1 = (1 - \tilde{x}_1^2 - \tilde{x}_2^2)\, \tilde{x}_1(t - \tau) - \tilde{x}_2(t - \tau) + \tilde{u}$$

$$\dot{\tilde{x}}_2 = \tilde{x}_1(t - \tau) \qquad\qquad 0 \le t \le T$$

The costate equations are,

$$\dot{\tilde{p}}_1 = 2\,\tilde{x}_1 + 2\,\tilde{x}_1\;\tilde{x}_1(t - \tau)\,\tilde{p}_1$$

$$\dot{\tilde{p}}_2 = 2\,\tilde{x}_2 + 2\,\tilde{x}_2\;\tilde{x}_1(t - \tau)\,\tilde{p}_1 \qquad\qquad T - \tau \le t \le T$$

$$\tilde{p}_1 = 2\,\tilde{x}_1 + 2\,\tilde{x}_1\;\tilde{x}_1(t - \tau)\,\tilde{p}_1 - (1 - \tilde{x}_1^2 - \tilde{x}_2^2)\,\tilde{p}_1 - \tilde{p}_2$$

$$\tilde{p}_2 = 2\,\tilde{x}_2 + 2\,\tilde{x}_2\;\tilde{x}_1(t - \tau)\,\tilde{p}_1 + \tilde{p}_1 \qquad\qquad 0 \le t \le T - \tau$$

The boundary conditions are,

$$\tilde{x}_1(t) = 10\,t$$

$$\tilde{x}_2(t) = 2 - 10\,t \qquad\qquad -\tau \le t \le 0$$

$$\tilde{p}(T) = 0$$

$$\tilde{p}(T - \tau_-) = \tilde{p}(T - \tau_+)$$

The optimal control $\tilde{u}^*(t,\tau)$ is given by,

$$\tilde{u}^*(t,\tau) = \frac{1}{2}\,\tilde{p}_1(t)\,.$$

The conjugate gradient method as developed in [8] was modified to solve the above time-lag equations. The near-optimal control is used as the initial guess for the control function in the optimum design.

Table 1 and Figure 1 show the values of the performance index obtained by all the three designs, for different values of time-lag, τ.

TABLE 1

Value of Time-Lag τ	Performance Index by Each Design		
	Optimal	Near-Optimal	Nominal
0.1	5.799	5.800	6.020
0.2	5.444	5.485	6.590
0.3	4.681	5.692	7.606

It can be seen that the performance index of the near-optimum design is close to that of optimal while the performance of nominal design is away from that of optimal. For time-lag $\tau = 0.2$, Figure 2 shows the controls obtained by the three designs. It can be noted that optimal and near-optimal controls are almost the same where as the nominal control is significantly different. Figures 3 and 4 show the values of the states $\tilde{x}_1$, $\tilde{x}_2$ for the three designs. Figure 5 shows the sensitivities $\tilde{\omega}_1$, $\tilde{\omega}_2$ and $\tilde{\upsilon}$.

CONCLUSIONS

The comparison of the responses and performances of the three designs on a nonlinear example indicates the closeness of the optimum and near-optimum designs where as the nominal design is away from the optimum. This can be expected because the nominal design does not take into account the history of the system during the time-lag period. The near-optimum design considers this history in obtaining the initial conditions for the state sensitivity and hence results in a better performance. In the near-optimum design, the solution of the sensitivity equations of Step 2 is straight forward as it does not involve any iterations. Thus the computational time required for near-optimum design is almost the same as that of the nominal design. This illustrates the computational simplicity and the practical applicability of the near-optimum design. The sensitivity functions developed in the near-optimum design clearly show the effect of time-lag on the system.

ACKNOWLEDGMENT

The authors are indebted to the National Science Foundation for financial support of this work under NSF Grant GK-4881.

REFERENCES

1. P. Kokotović and P. Sannuti, "Singular Perturbation Method for Reducing the Model Order in Optimal Control Design," IEEE Transactions on Automatic Control, Vol. 13, pp. 377-384, August 1968.

2. P. Sannuti and P. Kokotović, "Near Optimum Design of Linear Systems by Singular Perturbation Method," IEEE Transactions on Automatic Control, Vol. 14, pp. 15-22, February 1969.

3. P. Sannuti and P. Kokotović, "Singular Perturbation Method for Near Optimum Design of High-Order Nonlinear Systems," Proceedings of IFAC, Warsaw, Poland, 1969.

4. P. Sannuti, "Singular Perturbation Method in the Theory of Optimal Control," Ph.D. Thesis, Department of Electrical Engineering, University of Illinois, June 1968.

5. P. Sannuti, "Near Optimum Design of Time-Lag Systems by Singular Perturbation Method," Proceedings of JACC, 1970.

6. M. Athans and P. L. Falb, "Optimal Control," New York, McGraw-Hill Book Company, 1966.

7. A. P. Sage, "Optimum Systems Control," Englewood Cliffs, New Jersey, Prentice-Hall, Incorporated, 1968.

8. L. S. Lasdon, et al., "The Conjugate Gradient Method for Optimal Control Problems," IEEE Transactions on Automatic Control, Vol. 12, April 1967, pp. 132-138.

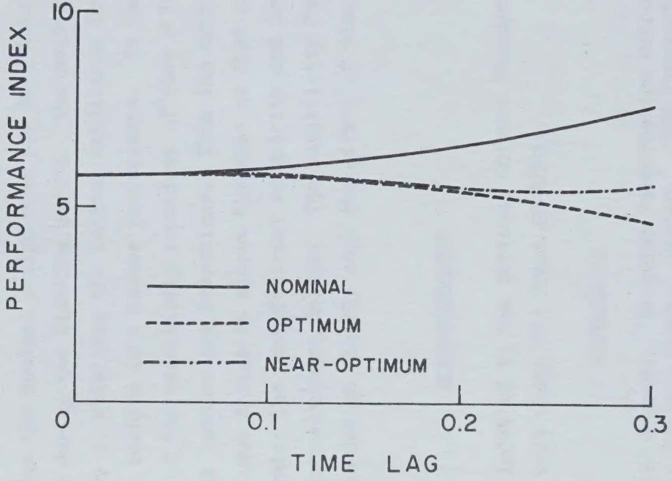


Figure 1. Performance index for the three designs

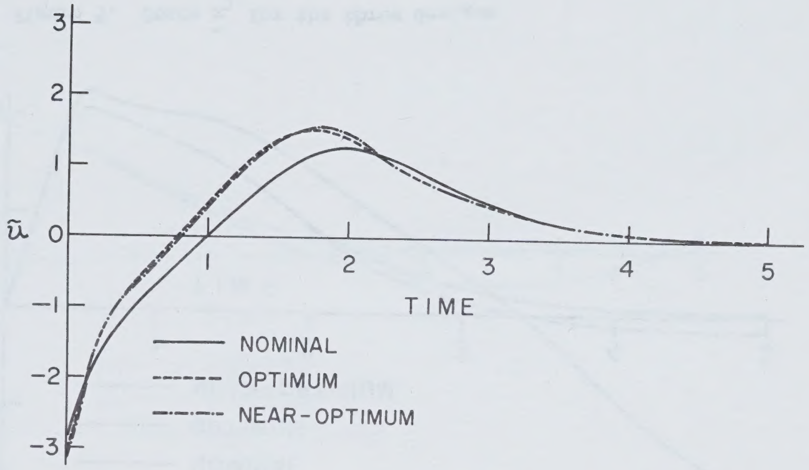


Figure 2. Control $\tilde{u}$ for the three designs

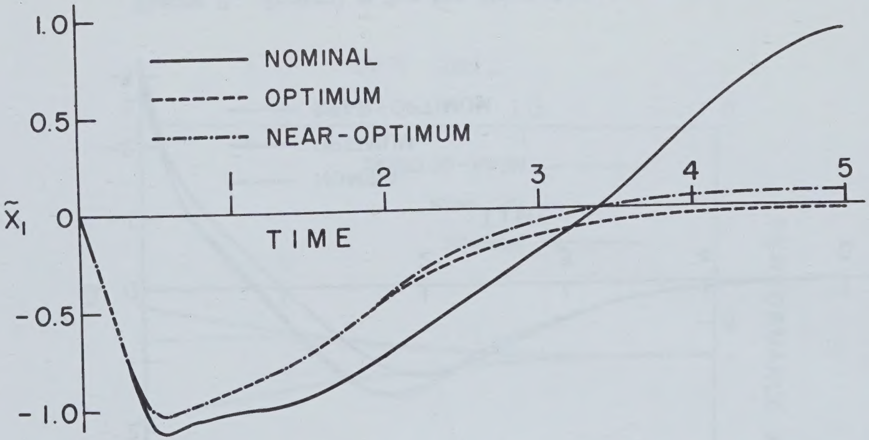


Figure 3. State $\tilde{x}_1$ for the three designs

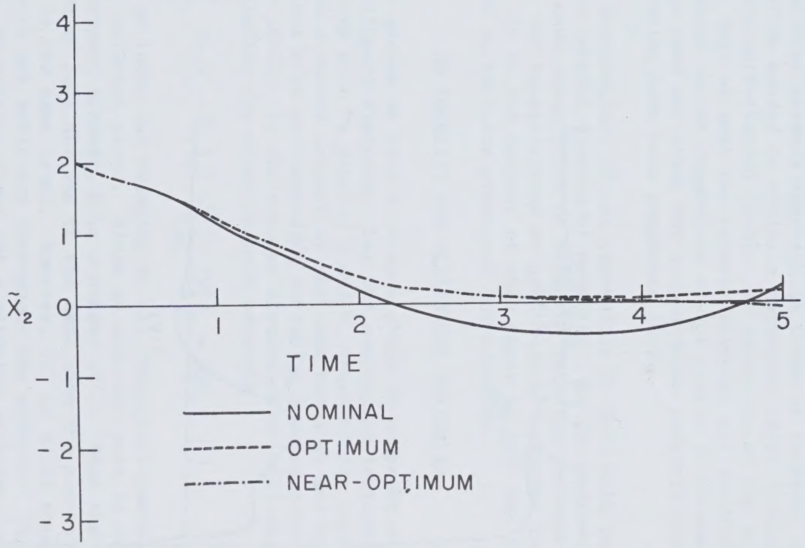


Figure 4. State $\tilde{x}_2$ for the three designs

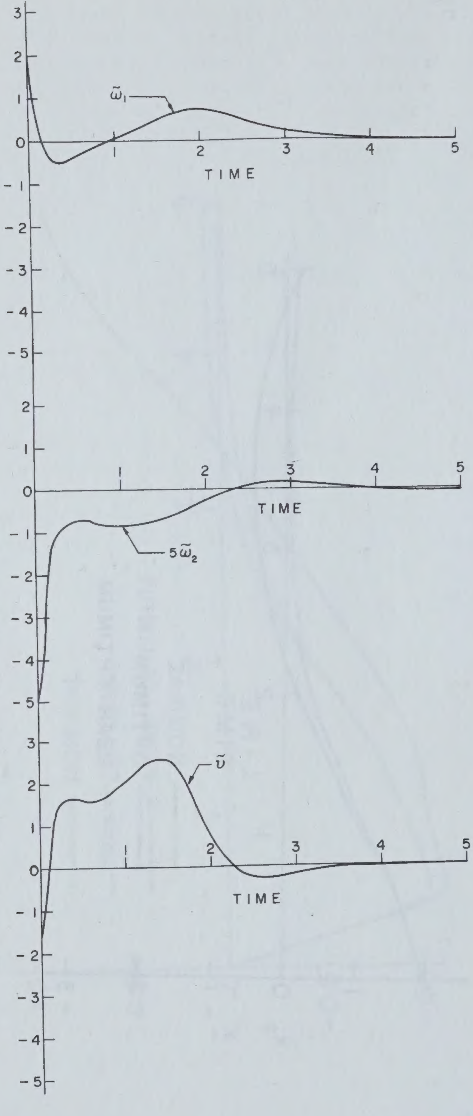


Figure 5. Sensitivity functions $\tilde{\omega}$, and $\tilde{\upsilon}$

SENSITIVITIES IN TERMS OF WAVE VARIABLES

JOHN W. BANDLER
Dept. of Electrical Engineering
McMaster University
Hamilton, Ontario, Canada

RUDOLPH E. SEVIORA
Dept. of Electrical Engineering
University of Toronto
Toronto, Ontario, Canada

Abstract It is the purpose of this paper to show how first- and second-order sensitivities and gradients with respect to network parameters can be evaluated directly in terms of wave variables. The concept of the adjoint network is employed. The application of these ideas to computer-aided network optimization is indicated.

INTRODUCTION

There is considerable interest and activity currently in the computation of sensitivities for lumped [1], [2] and distributed [3], [4] networks using the adjoint network approach in terms of currents and voltages. These are often needed to evaluate parameter space gradients in computer-aided network optimization [1-5]. An advantage of the adjoint network approach is that at most two network analyses are required for evaluation of the gradient vector regardless of the number of parameters. In certain problems the need may arise for second-order sensitivities and methods for their evaluation have been suggested [6-8].

It is inconvenient if not impossible to work with currents and voltages for certain classes of networks. In the microwave region, for example, a wave description of networks is often preferable. There is need, then, for sensitivities or gradients of networks in terms of wave variables. It is the purpose of this paper to show how they may be obtained and to indicate practical applications.

AN IDENTITY FOR SCATTERING VARIABLES

Let us assume we have a network which is composed of (in general) one-port and multiport elements. Let the normalized incident and reflected waves at every port be denoted a and b, respectively (see Figure 1(a)). Consider now a second network of the same topology and corresponding normalizations with corresponding variables denoted α and β, respectively (see Figure 1(b)). In the ensuing discussion we will need the following identity relating the waves in both networks

$$\sum_{i\varepsilon I} (b_i\alpha_i - a_i\beta_i) = \sum_{i\varepsilon E} (b_i\alpha_i - a_i\beta_i) \tag{1}$$

where I is an index set relating to all (interior) ports and E an index set relating to exterior ports. Since an exterior port is at the same time a port of a network element, E is a subset of I. That the above equation is true is more or less obvious in the case when the connected port pairs are normalized to the same level. However, (1) is valid even when the normalization levels are arbitrary throughout the network. This can be seen, for example, by writing first the scattering equation for an internal port junction (Figures 1(a) and (b))

This work was carried out under Grants A7239 and A5277 from the National Research Council of Canada.

$$\begin{bmatrix} a_p \\ a_q \end{bmatrix} = \underline{S}_j \begin{bmatrix} b_p \\ b_q \end{bmatrix} \tag{2a}$$

$$\begin{bmatrix} \alpha_p \\ \alpha_q \end{bmatrix} = \underline{S}_j \begin{bmatrix} \beta_p \\ \beta_q \end{bmatrix} \tag{2b}$$

where $\underline{S}_j$ is the scattering matrix of the junction, accounting for the change in normalization across the junction. Substituting (2a) and (2b) into the relevant terms of the left-hand side of (1) we have

$$\begin{bmatrix} b_p & b_q \end{bmatrix} \begin{bmatrix} \alpha_p \\ \alpha_q \end{bmatrix} - \begin{bmatrix} a_p & a_q \end{bmatrix} \begin{bmatrix} \beta_p \\ \beta_q \end{bmatrix}$$

$$= \begin{bmatrix} b_p & b_q \end{bmatrix} (\underline{S}_j - \underline{S}_j^T) \begin{bmatrix} \beta_p \\ \beta_q \end{bmatrix} . \tag{3}$$

Since the junction (normalization step) is reciprocal (3) is equal to zero, thus the left-hand side of (1) reduces to the sum over E. For an alternative proof using voltage and current concepts see Penfield et. al. [9].

Derivations in the following sections are based on a generalized version of (1) [9], namely

$$\sum_{i \varepsilon I} ((L\ b_i)\alpha_i - (L\ a_i)\beta_i) = \sum_{i \varepsilon E} ((L\ b_i)\alpha_i - (L\ a_i)\beta_i) \tag{4}$$

where L will be first-and second-order parameter space differential operators. Proof of (4) follows similar lines to the proof of (1).

FIRST-ORDER SENSITIVITIES

Suppose a parameter ϕ in the original network is varied without affecting topology. Further, let it be contained in an n-port subnetwork characterized by a scattering matrix $\underline{S}$ such that

$$\underline{b} = \underline{S}\underline{a} \tag{5}$$

where $\underline{b}$ and $\underline{a}$ are the appropriate n-element vectors. Then

$$\frac{\partial \underline{b}}{\partial \phi} = \frac{\partial \underline{S}}{\partial \phi} \underline{a} + \underline{S} \frac{\partial \underline{a}}{\partial \phi} . \tag{6}$$

The terms relating to this subnetwork on the left-hand side of (4) become

$$\underline{a}^T \left(\frac{\partial \underline{S}}{\partial \phi} \right)^T \underline{\alpha} + \left(\frac{\partial \underline{a}}{\partial \phi} \right)^T \left[\underline{S}^T \underline{\alpha} - \underline{\beta} \right] \tag{7}$$

which reduces to

$$\underline{a}^T \left(\frac{\partial \underline{S}}{\partial \phi}\right)^T \underline{\alpha} \tag{8}$$

if

$$\underline{\beta} = \underline{S}^T \underline{\alpha}. \tag{9}$$

Equation (9) defines the corresponding subnetwork of the auxiliary or, in current parlance, the adjoint network.

If all elements of the adjoint network are defined in this way

$$\sum_{i\varepsilon E} \left(\frac{\partial b_i}{\partial \phi} \alpha_i - \frac{\partial a_i}{\partial \phi} \beta_i \right) = \underline{a}^T \left(\frac{\partial \underline{S}}{\partial \phi} \right)^T \underline{\alpha} . \tag{10}$$

Then, by suitably terminating (and exciting) the network and its adjoint, the sensitivity of the desired wave can be computed. So, for example, if we are interested in the sensitivity of the ikth term of the overall scattering matrix, the network is analyzed for $a_k=1$, $a_j=0$ for all $j \varepsilon E$, $j \neq k$ (the latter is achieved by matched terminations), and the adjoint network excitations are zero except at the ith port, where a unit source is applied. Then (10) reduces to*

$$\frac{\partial S_{ik}}{\partial \phi} = \frac{\partial b_i}{\partial \phi} = \underline{a}^T \left(\frac{\partial \underline{S}}{\partial \phi} \right)^T \underline{\alpha} . \tag{11}$$

To evaluate the sensitivities one needs to know the partial derivatives of subnetwork scattering matrices. For the majority of common design components these can usually be found after some manipulation. As it happens, quite often $\partial \underline{S}/\partial \phi$ can be expressed in terms of the original $\underline{S}$ matrix which simplifies the evaluation of the sensitivities. Thus, for example, for a piece of waveguide of characteristic impedance Z using real normalization

$$\underline{a}^T \left(\frac{\partial \underline{S}}{\partial Z} \right)^T \underline{\alpha} = \frac{1}{2Z} \left(\underline{a}^T \underline{\alpha} - \underline{b}^T \underline{\beta} \right) . \tag{12}$$

Or, the arbitrariness in the normalization level can be used to advantage as demonstrated in the case of the sensitivity with respect to length of a piece of transmission line or waveguide. If the input and output waves are normalized on the impedance level equal to the characteristic impedance of the guide, then

$$\underline{a}^T \left(\frac{\partial \underline{S}}{\partial \ell} \right)^T \underline{\alpha} = -\gamma\, \underline{a}^T \underline{\beta} \tag{13}$$

where γ is the propagation constant of the waveguide.

---

*At the time of writing, it was brought to the authors' attention that some similar results have been independently derived [10].

SECOND-ORDER SENSITIVITIES

Consider two distinct non-topological parameters ϕ and ψ in the original network. Assume that they are contained in a subnetwork with scattering matrix $\underline{S}$. Applying $\partial/\partial\psi$ to (6) we get

$$\frac{\partial^2 \underline{b}}{\partial\psi\,\partial\phi} = \frac{\partial^2 \underline{S}}{\partial\psi\,\partial\phi}\,\underline{a} + \frac{\partial \underline{S}}{\partial\phi}\frac{\partial \underline{a}}{\partial\psi} + \frac{\partial \underline{S}}{\partial\psi}\frac{\partial \underline{a}}{\partial\phi} + \underline{S}\,\frac{\partial^2 \underline{a}}{\partial\psi\,\partial\phi}\ . \tag{14}$$

The terms relating to this subnetwork on the left-hand side of (4) now appropriately become

$$\underline{\alpha}^T\left(\frac{\partial^2 \underline{b}}{\partial\psi\,\partial\phi}\right) - \underline{\beta}^T\left(\frac{\partial^2 \underline{a}}{\partial\psi\,\partial\phi}\right)$$

$$= \underline{\alpha}^T\left(\frac{\partial^2 \underline{S}}{\partial\psi\,\partial\phi}\right)\underline{a} + \underline{\alpha}^T\frac{\partial \underline{S}}{\partial\phi}\frac{\partial \underline{a}}{\partial\psi} + \underline{\alpha}^T\frac{\partial \underline{S}}{\partial\psi}\frac{\partial \underline{a}}{\partial\phi} + \left(\underline{\alpha}^T\underline{S} - \underline{\beta}^T\right)\frac{\partial^2 \underline{a}}{\partial\psi\,\partial\phi}\ . \tag{15}$$

Using (9), we can reduce (15) to

$$\underline{\alpha}^T\left(\frac{\partial^2 \underline{S}}{\partial\psi\,\partial\phi}\right)\underline{a} + \underline{\alpha}^T\frac{\partial \underline{S}}{\partial\phi}\frac{\partial \underline{a}}{\partial\psi} + \underline{\alpha}^T\frac{\partial \underline{S}}{\partial\psi}\frac{\partial \underline{a}}{\partial\phi}\ . \tag{16}$$

If the whole auxiliary network is defined in accordance with (9) and if suitable terminations as described in the preceding section are applied, we obtain the counterpart of (11), namely,

$$\frac{\partial^2 S_{ik}}{\partial\psi\,\partial\phi} = \underline{\alpha}^T\left(\frac{\partial^2 \underline{S}}{\partial\psi\,\partial\phi}\right)\underline{a} + \underline{\alpha}^T\frac{\partial \underline{S}}{\partial\phi}\frac{\partial \underline{a}}{\partial\psi} + \underline{\alpha}^T\frac{\partial \underline{S}}{\partial\psi}\frac{\partial \underline{a}}{\partial\phi}\ . \tag{17}$$

A special case of interest is when ϕ and ψ are identical leading to

$$\frac{\partial^2 S_{ik}}{\partial\phi^2} = \underline{\alpha}^T\left(\frac{\partial^2 \underline{S}}{\partial\phi^2}\right)\underline{a} + 2\underline{\alpha}^T\frac{\partial \underline{S}}{\partial\phi}\frac{\partial \underline{a}}{\partial\phi}\ . \tag{18}$$

If ϕ belongs to a subnetwork with a scattering matrix $\underline{S}_\phi$ and ψ belongs to a subnetwork with a scattering matrix $\underline{S}_\psi$, it may readily be shown that

$$\frac{\partial^2 S_{ik}}{\partial\psi\,\partial\phi} = \underline{\alpha}_\phi^T\frac{\partial \underline{S}_\phi}{\partial\phi}\frac{\partial \underline{a}_\phi}{\partial\psi} + \underline{\alpha}_\psi^T\frac{\partial \underline{S}_\psi}{\partial\psi}\frac{\partial \underline{a}_\psi}{\partial\phi} \tag{19}$$

where subscripts ϕ and ψ denote quantities related to the appropriate subnetworks.

COMPUTATION OF PARTIAL DERIVATIVES

The results of the previous sections can be readily exploited in the evaluation of the gradients of practically any wave-based objective function with respect to a variety of parameters including frequency. The procedure is fairly well-known and is discussed in detail in a number of references [2-6], [11].

EXAMPLES

Convenient examples to test the ideas of this paper were provided by the resistively terminated cascade of transmission lines shown in Figure 2. Two-section and three-section 10:1 quarter-wave transformers optimum over 100% bandwidth were chosen. Their responses are also shown in Figure 2. The numerical values of the parameters of the two-section design were taken from reference [12] and of the three-section design from reference [13].

Tables I and II show the components of the gradient vector of the magnitude of the reflection coefficient ρ at 0.5 GHz estimated from 1% and .01% incremental changes in the parameters compared with those obtained from the results of this paper using one network analysis. Gradient calculations at other frequencies were also made by the present method. They were used by the first author in illustrating the necessary conditions for a minimax optimum in another paper presented at this conference [14].

CONCLUSIONS

This paper has shown how first- and second-order sensitivities and gradients with respect to network parameters can be evaluated directly in terms of wave variables, without recourse to voltages and currents. The method employs the concept of the adjoint network. As a result, the same benefits in terms of ease of implementation and computational efficiency as discussed by previous authors should be enjoyed. Efficient gradient methods of minimization can thus be employed in the optimal design by computer of networks for which the wave description is more natural or preferable.

ACKNOWLEDGEMENT

The authors would like to thank Dr. M. Sablatash of the Department of Electrical Engineering, University of Toronto, for his cooperation. The authors would also like to thank A. Lee-Chan of the McMaster Data Processing and Computing Centre for checking the numerical results given in this paper.

REFERENCES

[1] S.W. Director and R.A. Rohrer, "The generalized adjoint network and network sensitivities", IEEE Trans. Circuit Theory, vol. CT-16, pp. 318-323, August 1969.

[2] S.W. Director and R.A. Rohrer, "Automated network design - the frequency-domain case", IEEE Trans. Circuit Theory, vol. CT-16, pp. 330-337, August 1969.

[3] J.W. Bandler and R.E. Seviora, "Current trends in network optimization", IEEE Trans. Microwave Theory and Techniques, vol. MTT-18, December 1970.

[4] J.W. Bandler and R.E. Seviora, "Computation of sensitivities for noncommensurate networks", IEEE Trans. Circuit Theory, vol. CT-18, January 1971.

[5] R. Seviora, M. Sablatash and J.W. Bandler, "Least pth and minimax objectives for automated network design", Electronics Letters, vol. 6, pp. 14-15, January 1970.

[6] S.W. Director, "Increased efficiency of network sensitivity computations by means of LU factorization", Midwest Symposium on Circuit Theory, Austin, Texas, April 1969.

[7] P.J. Goddard and R. Spence, "Efficient method for the calculation of first- and second-order network sensitivities", Electronics Letters, vol. 5, pp. 351-352, August 1969.

[8] G.A. Richards, "Second-derivative sensitivity using the concept of the adjoint network", Electronics Letters, vol. 5, pp. 398-399, August 1969.

[9] P. Penfield, Jr., R. Spence and S. Duinker, "A generalized form of Tellegen's theorem", IEEE Trans. Circuit Theory, vol. CT-17, pp. 302-305, August 1970.

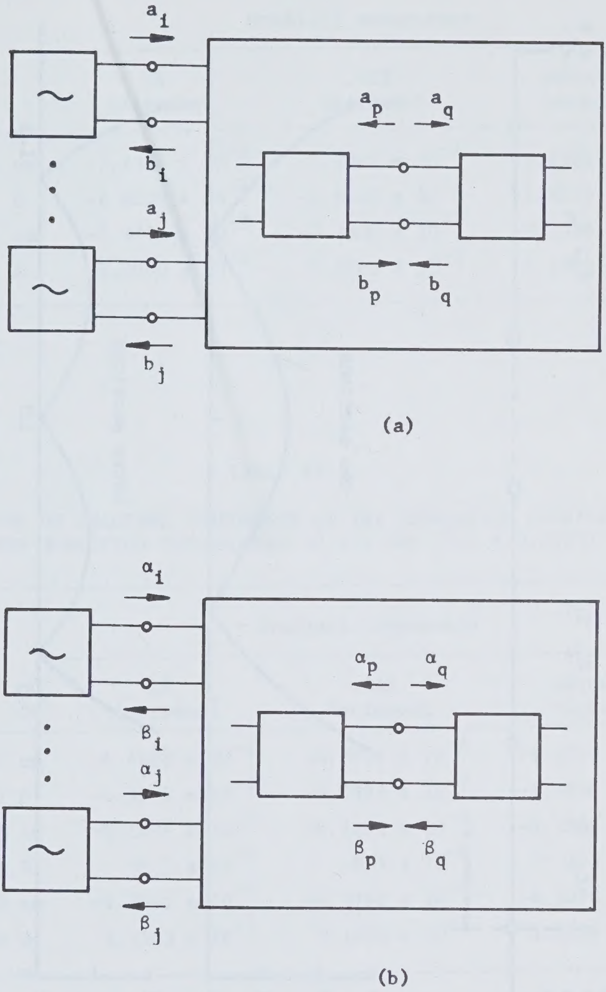
[10] V.A. Monaco and P. Tiberio, "On linear network scattering matrix sensitivity", Alta Frequenza, vol. 39, pp. 193-195, February 1970.

[11] G.C. Temes, "Exact computation of group delay and its sensitivities using adjoint-network concept", Electronics Letters, vol. 6, pp. 483-485, July 1970.

[12] J.W. Bandler and P.A. Macdonald, "Cascaded noncommensurate transmission-line networks as optimization problems", IEEE Trans. Circuit Theory, vol. CT-16, pp. 391-394, August 1969.

[13] J.W. Bandler and P.A. Macdonald, "Optimization of microwave networks by razor search", IEEE Trans. Microwave Theory and Techniques, vol. MTT-17, pp. 552-562, August 1969.

[14] J.W. Bandler, "Conditions for a minimax optimum", 8th Annual Allerton Conference on Circuit and System Theory, Urbana, Illinois, October 1970.



a_i
b_i
a_j
b_j
a_p
a_q
b_p
b_q
α_i
β_i
α_j
β_j
α_p
α_q
β_p
β_q

(a)

(b)

Figure 1. Two networks of the same topology and with corresponding port normalizations.

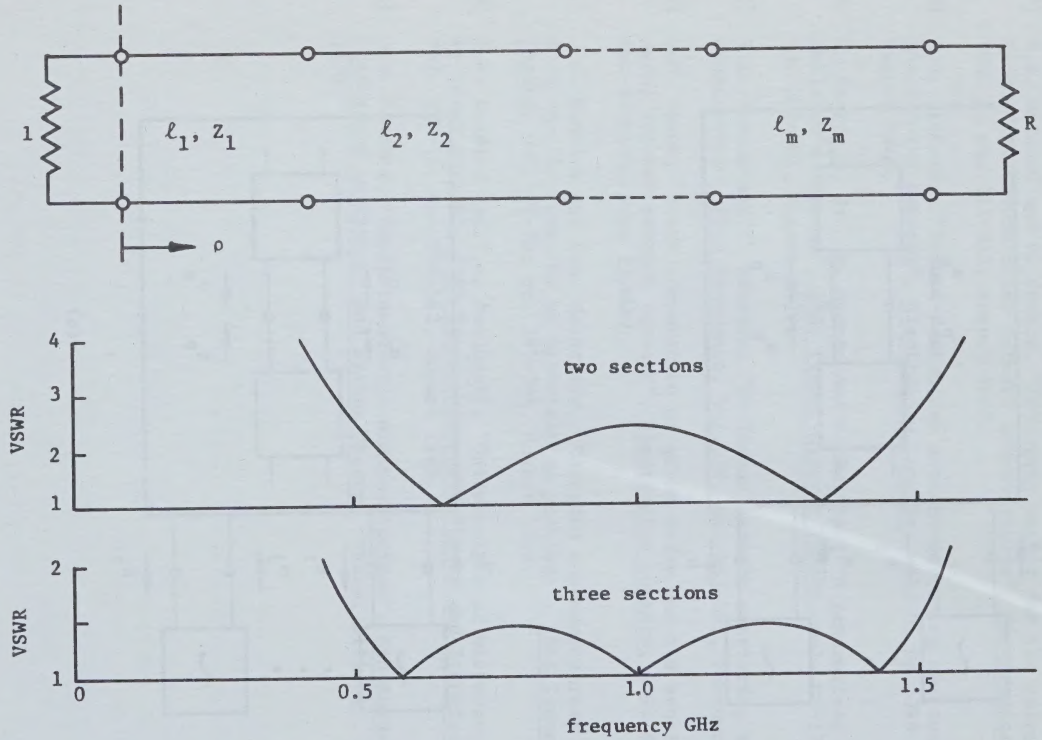

1
ℓ_1, Z_1
ℓ_2, Z_2
ℓ_m, Z_m
R
ρ
4
3
2
1
VSWR
two sections
2
1
VSWR
three sections
0
0.5
1.0
1.5
frequency GHz

Figure 2. m-section resistively terminated cascade of transmission lines. Also shown are the responses of 100% bandwidth optimum two- and three-section quarter-wave transformers for R=10.

TABLE I

COMPARISON OF GRADIENT COMPONENTS OF THE REFLECTION COEFFICIENT OF THE 2-SECTION TRANSFORMER AT 0.5 GHz ($|\rho| = 0.4286$)

Parameter values	Gradient components		
	1% increment	.01% increment	adjoint network
$\ell_1 = 7.49482$ cm	-7.4397×10^{-2}	-7.3337×10^{-2}	-7.3326×10^{-2}
$Z_1 = 2.2361\ \Omega$	-1.8250×10^{-1}	-1.8254×10^{-1}	-1.8254×10^{-1}
$\ell_2 = 7.49482$ cm	-7.3745×10^{-2}	-7.3330×10^{-2}	-7.3326×10^{-2}
$Z_2 = 4.4721\ \Omega$	9.0050×10^{-2}	9.1260×10^{-2}	9.1272×10^{-2}

TABLE II

COMPARISON OF GRADIENT COMPONENTS OF THE REFLECTION COEFFICIENT OF THE 3-SECTION TRANSFORMER AT 0.5 GHz ($|\rho| = 0.1973$)

Parameter values	Gradient components		
	1% increment	.01% increment	adjoint network
$\ell_1 = 7.49482$ cm	-4.4498×10^{-2}	-4.3777×10^{-2}	-4.3770×10^{-2}
$Z_1 = 1.63471\ \Omega$	-4.3461×10^{-1}	-4.3555×10^{-1}	-4.3556×10^{-1}
$\ell_2 = 7.49482$ cm	-9.1695×10^{-2}	-9.1294×10^{-2}	-9.1289×10^{-2}
$Z_2 = 3.16228\ \Omega$	-6.7×10^{-4}	-6.5×10^{-6}	4.0×10^{-7}
$\ell_3 = 7.49482$ cm	-4.3545×10^{-2}	-4.3767×10^{-2}	-4.3770×10^{-2}
$Z_3 = 6.11729\ \Omega$	1.1543×10^{-1}	1.1638×10^{-1}	1.1639×10^{-1}

SOLVING THE FELDTKELLER EQUATION

THOMAS G. MARSHALL, JR.
Rutgers University
New Brunswick, New Jersey

ABSTRACT

Two procedures are described for improving upon a solution to the Feldtkeller equation of two-port synthesis. A known polynomial is transformed in small increments into the solution polynomial by successive application of either procedure. In these procedures polynomials can be expressed as products of their factors, or alternatively, relative to any of a variety of bases, e.g. Lagrange polynomials. Squared and transformed frequency variables can be employed. The Euclidean algorithm is employed in the first method which avoids root finding. If the roots of the solution are desired, a second method employing a partial-fraction expansion effects the improvement. All distributions of reflection zeros can be accommodated in either procedure.

I. INTRODUCTION

The Feldtkeller equation is often solved by finding the roots (or linear and quadratic factors) of a polynomial. In the special case in which the desired solution is a Hurwitz polynomial, root finding can be avoided by using the method of Bauer which has also been considered in this application by MacWilliams [1]. The procedure described in Section III and V below also avoids root finding and has the advantage over Bauer's method of not requiring the solution to be a Hurwitz polynomial. It is therefore applicable to finding the numerator polynomial of a reflection coefficient with arbitrary distribution of reflection zeros between the left and right halves of the s-plane.

The procedures to be described are based upon two observations. The first is that an approximate solution to the Feldtkeller equation can be improved upon by the Euclidean algorithm. The partial-fraction method is just the special case of this improvement procedure obtained when the roots of the solution are desired.

The second observation is that approximate solutions to the Feldtkeller equation are always available if gain is regarded as a parameter which is varied from zero to the desired value. As the parameter is varied in steps, a sequence of polynomials is obtained, each varying only slightly from the preceding. The idea of regarding gain as a parameter was previously suggested [2] as a method of applying root-locus techniques [3] to the Feldtkeller equation. It was also observed that all polynomials could be expressed as products of factors in the root-locus method, so polynomials in powers of s could be avoided. A new filter synthesis procedure has been developed, the product method, in which polynomials are expressed as products of factors in all stages of the procedure [4]. The root-locus method and its reformulation below as the partial-fraction method, should both be applicable to solving the Feldtkeller equation in the product method.

The procedures to be described are also pertinent to the synthesis procedures employing matrix transformations [5,6], which like the product method, provide an alternative to expressing polynomials relative to the natural basis, $\left\{s^0, s^1, s^2, \ldots\right\}$. These transformation procedures are presently restricted to lowpass filters, but the solution methods described herein are general.

The solution methods to be described are applicable to the use of a squared or transformed variable as well as to the variable s [4,7]. It should be evident that various combinations of the preceding methods can be employed in solving the Feldtkeller equation, such as using a transformed variable and representing all polynomials in this variable relative to a set of Lagrange polynomials.

II. FORMULATION OF THE PROBLEM

A basic step in filter synthesis is the solution of the Feldtkeller equation [4,7],

$$H(s)H(-s) = 1 + K(s)K(-s), \qquad (2.1)$$

or the equivalent equation relating the scattering parameters,

$$S_{11}(s)S_{11}(-s) + S_{12}(s)S_{12}(-s) = 1. \qquad (2.2)$$

Let a and b be the monic polynomials having for their roots, respectively, the natural frequencies and the reflection zeros, and let z be a polynomial having as its roots the transmission zeros and normalized so that its lowest nonzero coefficient is unity. Then

$$S_{12}(s) = \sqrt{|\zeta_f|}\, z/a \qquad (2.3)$$

and

$$S_{11}(s) = \pm\, b/a \qquad (2.4)$$

where ζ_f is a real constant which determines the transducer gain. The following polynomial equation,

$$a(s)a(-s) = b(s)b(-s) + |\zeta_f| z(s)z(-s) \qquad (2.5)$$

is equivalent to (2.1) and (2.2) as can be seen by dividing (2.5) by $|\zeta_f| z(s)z(-s)$ to get (2.1) and dividing by $a(s)a(-s)$ to get (2.2) after transposing terms.

The problem is to factor $b(s)b(-s)$ to find b (alternatively, to factor $a(s)a(-s)$ to find a) when z and a (alternatively, z and b) are specified. A value for ζ_f must also be specified. Often, when b is to be determined, the value of ζ_f which maximizes transducer gain, say ζ_{max}, is desired, but this value is unknown at the outset. The following methods will give good approximations to ζ_{max} along with b, so the usual search for this value can be avoided or pursued after the approximation is available. When a is to be found, ζ_f is usually known from the stopband requirements, so no problem in its specification exists.

The parameter $\zeta^{(i)}$, which varies from zero to ζ_f, will now be introduced. For purposes of illustration assume that a is specified and b is to be found. Then the sequence of equations,

$$b^{(i)}(s)b^{(i)}(-s) = a(s)a(-s) - |\zeta^{(i)}| z(s)z(-s) \qquad (2.6)$$

is to be solved. Methods of doing this will be described in the subsequent sections.

III. IMPROVING UPON A SOLUTION WITHOUT ROOT-FINDING

A. Formulation

For convenience the following monic polynomials will be defined,

$$\bar{a}(s) = (-1)^n a(-s) \qquad (3.1)$$

and

$$\bar{b}(s) = (-1)^n b(-s), \qquad (3.2)$$

where n is the order of the filter and, hence, the degree of the polynomials in (3.1) and (3.2). In general, z will be of the form

$$z = s^m z_1 z_2 \ . \ . \ . \ z_\ell, \qquad (3.3)$$

where the factors defining the nonzero, finite transmission zeros would typically be of the form

$$z_i = (1 + s^2/\omega_i^2), \qquad (3.4)$$

although higher order factors could appear if complex conjugate transmission zeros are to be realized. In the lowpass case, $m = 0$.
Defining

$$\bar{z}(s) = (-1)^m z(-s) = z(s), \tag{3.5}$$

$$\zeta = |\zeta|(-1)^{m+n}, \tag{3.6}$$

and multiplying (2.6) by $(-1)^n$ yields

$$b\bar{b} = a\bar{a} - \zeta z^2, \tag{3.7}$$

in which a, $\bar{a}$, b, and $\bar{b}$ are all monic polynomials and the lowest nonzero coefficient of z and $\bar{z}$, the coefficient of s^m, is unity. Other normalizations for the polynomials can be employed, if desired, but the ones employed here are quite convenient. The superscript i's of Equation (2.6) are now omitted for convenience of notation.

To be explicit it will be assumed a is known and b is to be found. Assume at step i that an approximate solution, b^* is known to (3.7). Then an error e will be determined by

$$e = b\bar{b} - b^*\bar{b}^*. \tag{3.8}$$

In (3.8) the polynomial $b\bar{b}$ is known from (3.7), but only approximations b^* and $\bar{b}^*$ are known to its factors. It is desired to find an increment, Δb^*, such that

$$b' = b^* + \Delta b^* \tag{3.9}$$

is an improved solution, i.e. such that the new error

$$e' = b\bar{b} - b'\bar{b}' \tag{3.10}$$

is smaller than e.

To improve upon the solution, e will be cancelled to a first order approximation. To do this define increments Δb^* and $\Delta\bar{b}^*$ by

$$b^*\Delta\bar{b}^* + \bar{b}^*\Delta b^* = e. \tag{3.11}$$

Substituting (3.9) into (3.10) shows that the increments defined by (3.11) cancel e to a first order approximation, since the remaining error is

$$e' = -\Delta b^* \Delta\bar{b}^*. \tag{3.12}$$

The increments can be obtained from (3.11) by the Euclidean algorithm [8]. (References [5] and [9] discuss vector space methods for the Euclidean algorithm which provide flexibility in the mode of expressing the polynomials.) A sufficient condition for the solution to be obtainable is for e to be an even polynomial, which follows from (3.8), and for b^* and $\bar{b}^*$ to be relatively prime, which is normally true. The latter assumption permits the Euclidean algorithm to be performed with b^* and $\bar{b}^*$ to yield unique polynomials u and v with degrees less than n, satisfying

$$1 = b^*v + \bar{b}^*u. \tag{3.13}$$

Multiplying both sides of (3.13) by e shows that Δb^* and ue are equivalent modulo b^*. Then, by (3.8), Δb^* and $ub\bar{b}$ **are** equivalent modulo b^*. The increment Δb^* is the unique polynomial of lowest degree equivalent to ue, namely

$$\begin{aligned} \Delta b^* &= \text{remainder of } ue \text{ divided by } b^* \\ &= \text{remainder of } ub\bar{b} \text{ divided by } b^* . \end{aligned} \tag{3.14}$$

(It is seen that polynomials of degree $> n-1$ can be avoided when finding Δb^* if higher degree terms that appear in forming e $(b\bar{b})$ and ue $(ub\bar{b})$ are replaced with their equivalents modulo b^* that are of degree $< n$.)

In Section V, the solution procedure will be discussed further. It will then be evident that the evenness of e insures that $\Delta\bar{b}^*(s) = (-1)^n\Delta b^*(-s)$. This justifies the notation for $\Delta\bar{b}^*$ and insures that the improved polynomials b' and $\bar{b}'$ will also be similarly related, as the notation suggests. From these observations it is seen that $\Delta\bar{b}^*$ isn't needed since $\bar{b}'$ can be determined from b'.

The procedure, as described above, was programmed, and an example, summarized in Table I, was computed. A discussion of this example follows.

B. Example 1

A 3rd order Butterworth example will be considered for which the following information is known at the outset:

$$\begin{aligned} a &= (s+1)(s^2+s+1), \\ a\bar{a} &= s^6-1, \\ \zeta_{max} &= -1. \end{aligned} \tag{3.15}$$

It is specified that b is to have all its roots in the right half of the s plane and $\zeta_f = -63/64$. (This value of ζ_f corresponds to a termination ratio of 7/9. See Examples 2 and 3 of Reference [5].) In general the procedure is initiated with that solution, b, of (3.7) with $\zeta = 0$ which has the specified root distribution. Thus, in this example, the initial solution is

$$b' = b = (s-1)(s^2-s+1) = s^3 - 2s^2 + 2s-1 \tag{3.16}$$

The parameter ζ is to be equally incremented in the first ten steps.

In Step 1 in which $\zeta = -63/640$, the approximate solution, b^*, is the preceding b' of (3.16). Performing the Euclidean algorithm upon b^* and $\bar{b}^*$ yields for (3.13),

$$1 = b^*(-1/2 - 2s/3 - s^2/3) + \bar{b}^*(1/2 - 2s/3 + s^2/3) \tag{3.17}$$

It is seen that $\mathbf{v} = \bar{u}$. From (3.8)

$$e = (s^6 - 1 + 63/640) - (s^6 - 1) = 63/640. \tag{3.18}$$

In this case the remainder in (3.14) is ue itself, so

$$\Delta b^* = (63/640)(1/2 - 2s/3 + s^2/3). \tag{3.19}$$

The improved solutions, b', of (3.9) are indicated for this and the other steps in Table I. In the steps following Step 10, the desired solution of (3.7) with $\zeta = \zeta_f$ is successively improved.

TABLE I - Coefficients of Improved Solution Polynomial b'
with $\Delta\zeta = -63/640$ and $\zeta_f = -63/64$.

Step	ζ	s^0	s^1	s^2	s^3
0	0	-1.0000	2.0000	-2.0000	1.0
1	$\Delta\zeta$	-0.9508	1.9344	-1.9672	1.0
2	$2\,\Delta\zeta$	-0.8977	1.8619	-1.9301	1.0
3	$3\,\Delta\zeta$	-0.8413	1.7833	-1.8890	1.0
4	$4\,\Delta\zeta$	-0.7810	1.6972	-1.8429	1.0
5	$5\,\Delta\zeta$	-0.7156	1.6015	-1.7905	1.0
6	$6\,\Delta\zeta$	-0.6438	1.4932	-1.7292	1.0
7	$7\,\Delta\zeta$	-0.5634	1.3671	-1.6552	1.0
8	$8\,\Delta\zeta$	-0.4703	1.2142	-1.5612	1.0
9	$9\,\Delta\zeta$	-0.3564	1.0148	-1.4306	1.0
10	ζ_f	-0.2001	0.7135	-1.2140	1.0
11	ζ_f	-0.1391	0.5503	-1.0603	1.0
12	ζ_f	- .1257	0.5034	-1.0049	1.0
13	ζ_f	- .1250	0.5000	-1.0000	1.0
14	ζ_f	- .1250	0.5000	-1.0000	1.0
15	ζ_f	- .1250	0.5000	-1.0000	1.0

C. Example 2

This example illustrates a more general situation than Example 1 in which there are finite zeros of transmission, the reflection zeros are distributed in both halves of the s plane, and the Euclidean algorithm is required in its general form, that is a continued fraction expansion doesn't exist. The following information is known at the outset:

$$\begin{aligned} a &= (s+1)(s^2+s+1/4), \\ a\bar{a} &= s^6 - s^4 + s^2/4 - 1/4, \qquad (3.20) \\ z\bar{z} &= z^2 = (1+s^2/4)^2. \end{aligned}$$

The value of ζ_{max} can be seen to be between 0 and -1/4 from (3.7) at s = 0. Specify $\zeta^{(1)} = -1/8$. The distribution of reflection zeros is determined by specifying for $\zeta = 0$ in step 0 that

$$b' = b = (s+1)(s^2 - s + 1/2) = s^3 - s/2 + 1/2. \qquad (3.21)$$

In step 1, b^* is b' of step 0 and thus $b^*\bar{b}^* = a\bar{a}$. Equation (3.8) then gives

$$e = -\zeta^{(1)}z^2 = 1/8(1 + s^2/2 + s^4/16). \qquad (3.22)$$

The Euclidean algorithm has only 1 step in this case, as opposed to 3 steps for the usual case corresponding to a continued fraction expansion, which yields for (3.13)

$$1 = b^* - \bar{b}^*. \qquad (3.23)$$

Multiplying both sides of (3.23) by e and employing (3.14) yields

$$\Delta b^* = -17s^2/256 + s/256 - 1/8. \qquad (3.24)$$

IV. IMPROVING UPON THE ROOTS OF A SOLUTION BY PARTIAL-FRACTION EXPANSION

A. Formulation

Assume now that factors of $a\bar{a}$ and $z\bar{z}$ are known and those of $b\bar{b}$ are to be found. These products are all functions of s^2. The following procedure will be formulated in terms of the variable $x = s^2$, although it is equally applicable to the variable s. The following monic polynomial is defined

$$A(x) = a(s)\bar{a}(s), \qquad (4.1)$$

with similar definitions for B and Z. The procedure is also applicable if x is a transformed variable, but then the polynomials A, B, or Z may have some additional factors associated with the variable transformation in which case (4.1) should be modified. In this notation, (3.8) becomes

$$E = B - B^* = e, \tag{4.2}$$

and, assuming there are k factors of B, (3.9) becomes

$$B' = (B_1^* + \Delta B_1^*) \cdot \cdot \cdot (B_k^* + \Delta B_k^*). \tag{4.3}$$

As before, let first order terms equal E so that E' is of second order. The counterpart of (3.11) is then

$$\Delta B_1^*(B^*/B_1^*) + \cdot \cdot \cdot + \Delta B_k^*(B^*/B_k^*) = E, \tag{4.4}$$

in which terms in parenthesis are polynomials (associates of the Lagrange polynomials). Assume now the B_i^* are pairwise relatively prime.

The solution of (4.4) could be obtained by repeated application of the Euclidean algorithm, but another viewpoint is often simpler. Divide (4.4) by B* to obtain

$$\Delta B_1^*/B_1^* + \cdot \cdot \cdot + \Delta B_k^*/B_k^* = E/B^*. \tag{4.5}$$

The increments can thus be obtained by expanding E/B* in a partial fraction which is a very convenient procedure when the B_i^* are linear or quadratic. Assuming the factors are linear, Equation (4.5) together with (4.2) yields the equivalent of (3.14),

$$\Delta B_i^* = \text{Residue i of } E/B^* = \text{Residue i of } B/B^*. \tag{4.6}$$

This solution procedure can be seen to be essentially the same as the root-locus method if E is expanded in a Taylor series about each root of B*, although the viewpoint is different. In the formulation here, the loci are determined concurrently, which should be advantageous in checking for coincident roots.

This formulation can be conveniently expressed in terms of matrices by introducing the diagonal matrix $\underline{M}$, with entry ii the i^{th} root of B*, and the row vector $\underline{v}$, with entry i the $i^{\underline{\text{th}}}$ residue of 1/B*. The vector $\underline{v}B(\underline{M})$ then has the ΔB_i^* for its entries. This vector can be determined recursively directly from the factors of A and Z in (3.7), without explicitly forming the matrix $B(\underline{M})$.

When the roots of B are known, the roots of b can be found by taking square roots. These latter roots are the reflection zeros which are required for synthesis in many procedures. For example, the tridiagonal matrix $\underline{B}$ required in the methods of [5] and [6] can be directly formed from these roots. The partial-fraction procedure is well known, but the reader may wish to verify that the first step of Example 1, reformulated as in this section, yields $\Delta B_1^* = E/3 = |\zeta|/3$ and $\Delta B_{2,3}^* = (E/3)e^{\pm j120^\circ}$.

V. JUSTIFICATION AND EXTENSION OF THE PROCEDURES

A. Generalization to Include Common Factors

The following condition is necessary and sufficient for solutions to exist to equations of the form (3.11) and (4.4):

Condition 1 G.C.D. of $\{b^*; \bar{b}^*\}$ or $\{B^*/B_i^*\}$ divides e = E.

Condition 1 is satisfied in the normal step of these procedures with the greatest common divisor (G.C.D.) = 1, permitting (3.13) to be used. When a is to be determined and b and $\bar{b}$ have common factors, as for Chebyshev or Cauer filters, the following initialization procedure will insure that a^* and $\bar{a}^*$ in step 1 will have G.C.D. = 1. In step 0, let $\zeta = \zeta^{(1)} \neq 0$ in (3.7) and solve for (or guess) the roots of a^* and $\bar{a}^*$ by making a Taylor's series expansion of (3.7) about each root. For example, in the case of an inverted Chebyshev filter, the Taylor's series expansion at the multiple root of $b\bar{b}$ at s = 0 gives $a^*\bar{a}^* = s^{2n} - \zeta$, from which all the roots of a^* are easily obtained. The same procedure is applicable when the variable is x.

When b is the solution polynomial and transducer gain is to be maximized, b and $\bar{b}$ have a common factor. Accuracy will be lost as b^* and $\bar{b}^*$ approach b and $\bar{b}$, but the procedures give an approximation to the common factor, say b_1^*. This can be improved upon by the usual search along the $j\omega$ axis (or x axis), if desired, which will also give a better value of $\zeta_f = \zeta_{max}$. Condition 1 will then be satisfied because the common factor b_1^* divides e to the desired accuracy. The more general solution procedures of References [6] and [10] can then be employed to solve (3.11). Alternatively, replace 1 in (3.13) by b_1^* and e $(b\bar{b})$ in (3.14) by e/b_1^* $(b\bar{b}/b_1^*)$. The partial-fraction method affords the simple solution: if B_i^* and B_j^* are equal, specify $\Delta B_i^* = \Delta B_j^* = 0$, and calculate only the remaining ΔB_i^*.

B. Employing x as the Variable

In this section, let $b = b_e + sb_o$ where b_e and b_o are both even. Then, from (3.2) $\bar{b} = (-1)^n(b_e - sb_o)$. Consider the equation

$$by + \bar{b}x = e \tag{5.1}$$

in which degree b = degree $\bar{b}$ = n.

Property 1 If the G.C.D. of b and $\bar{b} = 1$, and if e is even of degree less than $2n$, then $x(s) = (-1)^n y(-s)$ for the unique x and y of degree $< n$.

By equating odd parts on both sides of (5.1) it is seen that

$$b_e(x_o + (-1)^n y_o) + b_o(x_e - (-1)^n y_e) = 0. \qquad (5.2)$$

However, the hypothesis implies that b_e and b_o are relatively prime. Thus, there is no nontrivial solution of (5.2), and the property is established. The solution methods described above, insure that e is always even, so the notation for the solution in (3.11) is warranted. If the even parts on both sides of (3.11) are equated, the following property results.

Property 2 $B^*_e \Delta B^*_e + (-xB^*_o)\Delta B^*_o = (-1)^n E/2.$ (5.3)

The notation $B^*_e(x) = b^*_e(s)$ and $B^*_o(x) = b^*_o(s)$ is employed. Equation (5.3) can be used as a solution procedure in the variable x, subject to the restriction stated in the following property, by employing the relation

$$B^* = (-1)^n (B_e^{*2} - xB_o^{*2}), \qquad (5.4)$$

and finding E from (4.2).

Property 3 If B^*_e and xB^*_o are relatively prime and

$$\deg B^*_e + \deg xB^*_o > \deg E, \qquad (5.5)$$

then (5.3) and (3.11) **give** the same **increment Δb^*.** Condition (5.5) is always satisfied if the sum of the roots of b^* is nonzero.

In Example 2, (5.5) is not satisfied. In spite of the existence of these exceptional cases, (5.3) is attractive because the given polynomials have only one half as many terms as those of (3.11). In the exceptional case, E could be truncated to satisfy (5.5) or (3.11) used.

C. Example 3

The simplification in employing (5.3) rather than (3.11) is illustrated by considering Example 1. From (3.16), $B^*_e = -2x - 1$ and $(-xB^*_o) = -x^2 - 2x$. The Euclidean algorithm gives

$$1 = B^*_e(-2x/3 - 1) + (-xB^*_o)(4/3). \qquad (5.6)$$

Multiplication of (5.6) by $(-1)^3 E/2$, with E obtained from (3.18), **gives (3.19).**

D. Application to Polynomial Factoring

Given a polynomial a to be factored, choose z arbitrarily, say $z = 1$. Let $b^* = \bar{a}$ for step 1 and solve (3.7) for values of ζ from 0 to a value, $\zeta_1 \simeq \zeta_{max}$, at which a common factor, b_1, of b^* and $\bar{b}^*$ is placed in evidence by the Euclidean algorithm. Example 1 is relevant thus far. At this point, the procedure will be continued by letting $|\zeta|$ decrease from $|\zeta_1|$ to zero. The new starting b^* and $\bar{b}^*$ are specified so as to interchange the factor which at $\zeta = \zeta_1$ was b_1. Thus, the final b*, b_f, for $\zeta = 0$ will have the interchanged factor in common with a, and hence the Euclidean algorithm applied to a and b_f will identify this factor. After removal of the factor, the process can be repeated.

VI. ACKNOWLEDGMENT

The author wishes to thank Dr. Skwirzynski for discussing his work with him and providing him a preliminary version of his paper. He also wishes to thank Mr. Louis Rubinfield for his assistance in programming and testing the procedures.

VII. BIBLIOGRAPHY

[1] F. Jessie MacWilliams, "An iterative method for the direct Hurwitz-factorization of a polynomial," IEEE Trans. Circuit Theory, vol. CT-5, pp. 347-352, Dec. 1958.

[2] T. G. Marshall, Jr., Ladder Network Synthesis by Similarity Transformation. Princeton, N.J.:RCA, Oct. 1965. Preprint of paper for 3rd Allerton Conf.

[3] R. W. Klopfenstein, "Zeros of nonlinear functions," Journal A.C.M., vol. 8-3, pp. 366-373, July 1961.

[4] J. K. Skwirzynski, "On synthesis of filters," to appear in IEEE Trans. Circuit Theory.

[5] T. G. Marshall, Jr., "Synthesis of RLC ladder networks by matrix tri-diagonalization," ibid, vol. CT-16, pp. 39-46, Feb. 1969.

[6] ——————————, "Analysis and synthesis of filters by transformation of matrix binomials," Proc. 1969 Allerton Conf., pp. 762-772.

[7] H.J. Orchard and G.C.Temes, "Filter design using transformed variables," IEEE Trans. Circuit Theory, vol. CT-15, pp. 385-408, Dec. 1968.

[8] M. Bôcher, Introduction to Higher Algebra, New York: MacMillan, 1907.

[9] T. G. Marshall, Jr., "A greatest common divisor algorithm for a vector space," Proc. 1968 Allerton Conf., pp. 374-385.

[10] R. J. Levit, "A minimum solution of a diophantine equation," Amer. Math. Mon., vol. 63, pp. 647-651, 1956.

NECESSARY AND SUFFICIENT CONDITIONS FOR THE COMPACTNESS OF AN RC TWO-PORT NETWORK*

M. J. MASSARO **
IIT Research Institute
Annapolis, Maryland

S. D. BEDROSIAN
Moore School of Electrical Engineering
University of Pennsylvania
Philadelphia, Pa.

ABSTRACT

Necessary and sufficient conditions are given for the compactness of the open-circuit impedance or short-circuit admittance matrix of an RC two-port network when the specified short-circuit admittance or open-circuit impedance matrix of the network is compact.

INTRODUCTION

The conditions for the compactness of the short-circuit admittance or open-circuit impedance matrix of an RC two-port network are discussed extensively in the literature [1]-[5]. If a given set of impedance (admittance) parameters is compact, there is no guarantee that the corresponding set of admittance (impedance) parameters will be compact. A necessary condition for a set of impedance (admittance) parameters to be compact when the corresponding specified set of admittance (impedance) parameters is compact was given by Slepian [4] for symmetrical and unsymmetrical RC two-ports. For symmetrical RC two-ports, Balabanian [1] gave the necessary and sufficient conditions for a set of impedance (admittance) parameters to be compact when the corresponding specified set of admittance (impedance) parameters is compact.

This paper presents the necessary and sufficient conditions for a set of impedance (admittance) parameters to be compact when the admittance (impedance) parameters are compact. The results are applicable to both symmetrical and unsymmetrical RC two-port networks.

SPECIAL PROPERTIES OF RC TWO-PORT NETWORKS

Certain properties of RC two-port network functions will be employed in the proofs of the theorems given in a following section. Some of these properties are summarized in this section.

The input parameters y_{11}, y_{22}, z_{22}, are RC driving point functions. All of their poles and zeros are simple and lie on the negative real axis. The residues at all of the $-\sigma$ axis poles must be positive and real. The first critical frequency of a driving point admittance function is a zero; whereas, the first critical frequency of a driving point impedance function is a pole.

---

*This paper is based on part of the results of the doctoral dissertation "Synthesis of a Class of Immittance Matrices by RC Three-Terminal Two-Ports," submitted by M. J. Massaro to the faculty of the Graduate School of Arts and Sciences, University of Pennsylvania, 1969, under the supervision of Dr. S. D. Bedrosian.

**Dr. Massaro was formerly employed by John L. Wildermuth & Associates, Rockville, Md.

The admittance or y parameters of a network can be expressed by a partial fraction expansion as follows:

$$(\pm)\ y_{jp} = \lambda_{jp}^{(0)} + s\lambda_{jp}^{(\infty)} + \sum_{i=1}^{n-1} \frac{s\lambda_{jp}^{(i)}}{(s+\sigma_i)} \qquad (1)$$

$$j = 1,2 \text{ and } p = 1,2$$

where the positive sign holds for $j=p$ and the negative sign for $j\neq p$, $-\sigma_i$, is a $-\sigma$ axis pole of the y parameters, $\lambda_{jp}^{(i)}$ is the residue of $(\pm)\ \frac{y_{jp}}{s}$ at the pole $-\sigma_i$, $k_{jp}^{(0)}$ is the residue at zero, and $k_{jp}^{(\infty)}$ is the residue of $\pm\ \frac{y_{jp}}{s}$ at the pole at infinity.

Similarly, the z parameters can be expressed by partial fractions as

$$z_{jp} = k_{jp}^{(\infty)} + \frac{k_{jp}^{(0)}}{s} + \sum_{i=1}^{n-1} \frac{k_{jp}^{(i)}}{(s+\nu_i)} \qquad (2)$$

$$j=1,2 \text{ and } p=1,2$$

where $-\nu_i$ is a $-\sigma$ axis pole of the z parameters, $k_{jp}^{(i)}$ is the residue of z_{jp} at the pole $-\nu_i$, $k_{jp}^{(\infty)}$ is the residue at infinity, $k_{jp}^{(0)}$ is the residue of z_{jp} at the pole occuring at the origin. All of the residues of the admittance and impedance parameters must satisfy the residue condition which can be expressed as

$$k_{11}\, k_{22} - k_{12}^2 \ \geq 0 \qquad (3)$$

and

$$\lambda_{11}\, \lambda_{22} - \lambda_{12}^2 \ \geq 0 \qquad (4)$$

respectively.

DEFINITIONS OF COMPACTNESS

There are two widely accepted definitions of compactness, one by Slepian [4] and the other by Dasher [2]. Slepian's definition can be stated as follows:

> A set of admittance or impedance parameters of an RC two-port network is compact if the residue condition is satisfied with an equality sign at each of the poles of the set of parameters.

Dasher's definition can be stated as follows:

> The impedance and admittance parameters of an RC two-port network are compact if all of the residues of both parameter sets satisfy the residue condition with an equality sign.

Note that Slepian's definition involves only one of the parameter sets of a network and only the residues at the $-\sigma$ axis poles of the parameters are considered. On the other hand, Dasher includes both parameter sets and all of the residues of the parameter sets are considered. In this paper private poles will be considered to be compact, that is, $k_{11}=0$ or $k_{22}=0$ and $k_{12}=0$, hence $k_{11}k_{22}-k_{12}^2=0$. But semi-private poles will not be considered to be compact, that is, $k_{12}=0$ and hence $k_{11}k_{22}-k_{12}^2\neq 0$. The corresponding results apply to admittance parameters.

NECESSARY AND SUFFICIENT CONDITIONS FOR COMPACTNESS

The following theorem gives the necessary and sufficient conditions for the impedance (admittance) parameters of an RC two-port network to satisfy Slepian's definition of compactness when the corresponding specified admittance (impedance) parameters of the network are compact.

THEOREM 1:

If the residues of a set of impedance (admittance) parameters are compact at each of the $-\sigma$ axis poles, then the residues of the admittance (impedance) parameters are compact at each of the $-\sigma$ axis poles if and only if the impedance parameters do not have a common zero.

Proof of Sufficiency:

In the proofs to follow it is assumed that a compact set of impedance parameters is specified. A similar proof can be shown to hold if a compact set of admittance parameters is specified instead. The following set of z parameters has compact residues at all $-\sigma$ axis poles, the polynomials have been reduced to lowest terms, all private and semi-private poles have been removed, and the parameters do not share a common zero:

$$z_{11}=\frac{n_{11}(s)}{d(s)}=\frac{a_n s^n+\ldots+a_o}{d_n s^n+\ldots+d_o} \tag{5a}$$

$$z_{22}=\frac{n_{22}(s)}{d(s)}=\frac{b_n s^n+\ldots+b_o}{d_n s^n+\ldots+d_o} \tag{5b}$$

$$z_{12}=\frac{n_{12}(s)}{d(s)}=\frac{c_m s^m+\ldots+c_o}{d_n s^n+\ldots+d_o} \tag{5c}$$

where $o \leq m \leq n$. Usually, d (s) can be of degree n or n + 1, but if d (s) is of degree n + 1, the residues at infinity of the parameters will be zero and the parameters will share a common zero at infinity. Thus, d(s) must be of degree n. The residues at the $-\sigma$ axis poles can be expressed in terms of the numerator and denominator polynomials as

$$k_{22}^{(i)}=\frac{n_{22}(-\nu_i)}{d'(-\nu_i)} \qquad k_{12}^{(i)}=\frac{n_{12}(-\nu_i)}{d'(-\nu_i)} \qquad k_{11}^{(i)}=\frac{n_{11}(-\nu_i)}{d'(-\nu_i)} \tag{6}$$

where $\dfrac{d'(s)}{(s+\nu_i)}=d(s)$ and $d'(-\nu_i)\neq 0$

since the poles of the parameters must be simple. From (3) and (6) we have

$$\frac{n_{11}(-\nu_i)n_{22}(-\nu_1)-(n_{12}(-\nu_1))^2}{(d'(-\nu_1))^2}=0 \tag{7}$$

for i = 1,2, •••, n; thus the numerator polynomial of the left side of (7) is zero at the frequency $-\nu_1$. The polynomial $[(n_{11}(s)n_{22}(s)-(n_{12}(s))^2]$ contains all of the zeros of d(s) and d(s) divides this polynomial.

The y parameters can be expressed as [6]

$$y_{11}=\frac{n_{22}(s)}{e(s)} \qquad y_{22}=\frac{n_{11}(s)}{e(s)} \qquad -y_{12}=\frac{n_{12}(s)}{e(s)} \tag{8}$$

where;

$$e(s)=\frac{n_{11}(s)n_{22}(s)-(n_{12}(s))^2}{d(s)}\,. \tag{9}$$

The poles of the admittance parameters must be simple. Thus,if e(s) has a double zero at some frequency, $n_{11}(s)$, $n_{22}(s)$, and $n_{12}(s)$ must have a zero at this frequency also so that cancellations can occur. If e(s) has a triple zero, $n_{11}(s)$, $n_{22}(s)$, and $n_{12}(s)$ must have a double zero at this frequency for the proper number of cancellations to occur. This is not possible since the numerator polynomials of the specified z parameters must have simple zeros. Thus, e(s) cannot have a zero of the third order or higher. For the persent case, the numerator polynomials of the parameters do not have a common zero. Hence, the zeros of e(s) must be simple.

The polynomial e(s) will be of degree 2n-n or n if the residues at infinity of the z parameters are not compact and e(s) will be of degree 2n-1-n or n-1 if the residues at infinity are compact. This can be shown as follows. The residues of the impedance parameters at infinity can be expressed in terms of the coefficients of the polynomials as

$$k_{11}^{(\infty)}=\frac{a_n}{d_n} \qquad k_{22}^{(\infty)}=\frac{b_n}{d_n} \qquad k_{12}^{(\infty)}=\frac{c_n}{d_n} \tag{10}$$

where it has been assumed that m = n but $k_{12}^{(\infty)}=0$ for m < n. If the residues at infinity are not compact, then from (10) we have

$$\frac{a_n b_n-c_n^2}{d_n^2}\neq 0. \tag{11}$$

Thus, the polynomial $[n_{11}(s)n_{22}(s)-(n_{12}(s))^2]$ will be of degree 2n, e(s) will be of degree n, and the y parameters will not have a pole at infinity. If the residues at infinity are compact, then from (10)

$$\frac{a_n b_n-c_n^2}{d_n^2}=0. \tag{12}$$

Thus, $[n_{11}(s)n_{22}(s)-(n_{12}(s))^2]$ will be of degree (2n-1) and e(s) will be of degree n-1. It can be seen from (8) that the y parameters will have a pole at infinity. If e(s) can be expressed as

$$e(s)=e_{n-1}s^{n-1}+\cdots+e_o \tag{13}$$

then from (8) and (5) the residues of the y parameters at the pole at infinity are

$$\lambda_{11}^{(\infty)} = \frac{b_n}{e_{n-1}} \qquad \lambda_{22}^{(\infty)} = \frac{a_n}{e_{n-1}} \qquad \lambda_{12}^{(\infty)} = \frac{c_n}{e_{n-1}}. \tag{14}$$

From (12) and (14) we have

$$\lambda_{11}^{(\infty)} \lambda_{22}^{(\infty)} - (\lambda_{12}^{(\infty)})^2 = \frac{b_n a_n - c_n^2}{e_{n-1}^2} = 0 \tag{15}$$

and the residues of the pole at infinity of the y parameters are compact.

The zeros of e(s), excluding the ones cancelled by d(s), are the zeros of $[n_{11}(s)n_{22}(s)-(n_{12}(s))^2]$. There are two possible cases to be considered:

1. $[n_{11}(-\sigma_1)n_{22}(-\sigma_1)-(n_{12}(-\sigma_1))^2] = 0$
2. $[n_{11}(-\sigma_1)n_{22}(-\sigma_1)] = 0$ and $n_{12}(-\sigma_1) = 0$.

Case 1:

Using (8), the residues of the y parameters at a finite pole $-\sigma_1$ can be expressed as

$$\lambda_{11}^{(1)} = \frac{n_{22}(-\sigma_1)}{(-\sigma_1)e'(-\sigma_1)} \qquad \lambda_{22}^{(1)} = \frac{n_{11}(-\sigma_1)}{(-\sigma_1)e'(-\sigma_1)}$$

$$\lambda_{12}^{(1)} = \frac{n_{12}(-\sigma_1)}{(-\sigma_1)e'(-\sigma_1)} \tag{16}$$

and

$$\lambda_{11}^{(1)} \lambda_{22}^{(1)} - (\lambda_{12}^{(1)})^2 = \frac{n_{11}(-\sigma_1)n_{22}(-\sigma_1)-(n_{12}(-\sigma_1))^2}{(-\sigma_1)^2(e'(-\sigma_1))^2} \tag{17}$$

where $e'(s) = \frac{e(s)}{(s+\sigma_1)}$, $-\sigma_1 \neq 0$ since the admittance parameters of an RC network cannot have a pole at the origin, and $e'(-\sigma_1) \neq 0$ since the zeros of e(s) must be simple. From (9) we have

$$e(-\sigma_1) = 0 = \frac{n_{11}(-\sigma_1)n_{22}(-\sigma_1)-(n_{12}(-\sigma_1))^2}{d(-\sigma_1)} \tag{18}$$

where $d(-\sigma_1)$ is finite. Thus, from (17) and (18) we have

$$\lambda_{11}^{(1)} \lambda_{22}^{(1)} - (\lambda_{12}^{(1)})^2 = 0 \tag{19}$$

and the residues of the y parameters at the pole $-\sigma_1$ are compact.

Case 2:

If $n_{11}(-\sigma_1) = 0 = n_{12}(-\sigma_1)$; then $e(-\sigma_1) = 0$ and the y parameters can be expressed as

$$y_{11} = \frac{n_{22}(s)}{e'(s)(s+\sigma_1)} \qquad y_{22} = \frac{n'_{11}(s)(s+\sigma_1)}{e'(s)(s+\sigma_1)} = \frac{n'_{11}(s)}{e'(s)}$$

$$-y_{12} = \frac{n'_{12}(s)(s+\sigma_1)}{e'(s)(s+\sigma_1)} = \frac{n'_{12}(s)}{e'(s)} \tag{20}$$

where the prime indicates that the factor $(s+\sigma_1)$ has been removed. Thus, y_{11} has a private pole at frequency $-\sigma_1$ and the residues of the y parameters at the pole $-\sigma_1$ are compact.

Similarly, if $n_{22}(s)$ and $n_{12}(s)$ have a common zero it can be shown that y_{22} will have a private pole.

Proof of Necessity:

Let us assume that a set of impedance parameters and the corresponding set of admittance parameters have compact residues at all $-\sigma$ axis poles and the impedance parameters share a common zero. There are two possible cases to be considered:

1. The numerator polynomials of the impedance parameters share a a common zero at frequency $-\gamma$; thus,
   $n_{11}(-\gamma) = n_{22}(-\gamma) = n_{12}(-\gamma) = 0$
2. The residues at infinity are zero; thus,
   $k_{11}^{(\infty)} = k_{22}^{(\infty)} = k_{12}^{(\infty)} = 0.$

Case 1:

Since the impedance parameters have compact residues at all $-\sigma$ axis poles, $d(s)$ divides $[n_{11}(s)n_{22}(s)-n_{12}^2(s)]$ and $e(s)$ is a polynomial. The polynomial $e(s)$ can be written as

$$e(s) = (s+\gamma)^2 \frac{[n'_{11}(s)n'_{22}(s)-(n'_{12}(s))^2]}{d(s)} \tag{21}$$

where the factor $(s+\gamma)$ has been removed from all the primed polynomials. Thus, $e(s)$ has a double zero at $-\gamma$. Forming the admittance parameters cancellations occur as follows:

$$y_{11} = \frac{n'_{22}(s)}{e''(s)(s+\gamma)} \qquad y_{22} = \frac{n'_{11}(s)}{e''(s)(s+\gamma)} \qquad -y_{12} = \frac{n'_{12}(s)}{e''(s)(s+\gamma)} \tag{22}$$

where

$$e''(s) = \frac{e(s)}{(s+\gamma)^2} \tag{23}$$

The residues at the pole $-\gamma$ are

$$\lambda_{11}^{(1)} = \frac{n'_{22}(-\gamma)}{(-\gamma)e''(-\gamma)} \qquad \lambda_{22}^{(1)} = \frac{n'_{11}(-\gamma)}{(-\gamma)e''(-\gamma)} \qquad \lambda_{12}^{(1)} = \frac{n'_{12}(-\gamma)}{(-\gamma)e''(-\gamma)} \tag{24}$$

and
$$\lambda_{11}^{(1)}\lambda_{22}^{(1)} - (\lambda_{12}^{(1)})^2 = \frac{n'_{11}(-\gamma)n'_{22}(-\gamma)-(n'_{12}(-\gamma))^2}{(-\gamma)^2(e''(-\gamma))^2}. \tag{25}$$

The residues at the pole $-\gamma$ are compact only if the numerator of the right side of (25) is zero. If the residues at the pole $-\gamma$ are compact, then,

$$\lim_{s\to -\gamma} e''(s) = \lim_{s\to -\gamma} \frac{[n'_{11}(s)n'_{22}(s)-(n'_{12}(s))^2]}{d(s)} = 0. \tag{26}$$

But, if $e''(-\gamma) = 0$, then $e(s)$ has a triple zero at $-\gamma$ and from previous discussions this is not possible. The y parameters have a non-compact pole at $-\gamma$. Thus, the admittance parameters cannot have compact residues at all $-\sigma$ axis poles if the impedance parameters have a common zero. Similarly, it can be shown that the impedance parameters will not be compact at all $-\sigma$ axis poles if the admittance parameters have a common zero.

Case 2:

If $d(s)$ is of degree $n+1$, then $n_{11}(s)$ and $n_{22}(s)$ must be of degree n and $n_{12}(s)$ can be, at most, of degree n. It can be seen from (9) that the degree of $e(s)$ is $n-1$ and $e(s)$ is a polynomial since the impedance parameters are compact at all $-\sigma$ axis poles and $d(s)$ divides the numerator term. The polynomial $e(s)$ also will have the same form as (13). The admittance parameters given by (8) will have a pole at infinity. The residues of the admittance parameters at this pole can be expressed by (14) and the residues will not be compact unless (15) holds. But if $a_n b_n - c_n^2=0$, the polynomial $[n_{11}(s)n_{22}(s)-(n_{12}(s))^2]$ will be of degree $2n-1$ and $e(s)$ of degree $n-2$. Thus, the admittance parameters will have a double pole at infinity. The admittance parameters must be a valid set of parameters and cannot possess a double pole; therefore, the admittance parameters do not have a compact pole at infinity. The original assumption is contradicted; impedance parameters cannot have a common zero at infinity otherwise, the admittance parameters will have a non-compact pole. This completes the proof of Theorem 1.

The following theorem gives the necessary and sufficient conditions for a set of impedance (admittance) parameters to satisfy Dasher's definition of compactness if the corresponding set of admittance (impedance) parameters are compact.

Theorem 2:

If all of the residues of a set of admittance (impedance) parameters of an RC two-port network are compact, then all of the residues of the impedance (admittance) parameters are compact if and only if the admittance (impedance) parameters do not have a common zero and the admittance (impedance) parameters have a pole at infinity (pole at the origin).

The proof showing that the residues of the impedance (admittance) parameters at all of the $-\sigma$ axis poles are compact is similar to the proof of Theorem 1. It remains to be shown that the existence of a pole at infinity (pole at the origin) for the admittance (impedance) parameters is an additional necessary and sufficient condition for the residues at infinity (residues at zero) of the impedance (admittance) parameters to be compact.

Proof of Sufficiency:

Let us assume that we are given a set of admittance parameters which has compact residues, a pole at infinity, and no common zeros. It was shown previously that if the residues of the impedance parameters at infinity were compact, the admittance parameters will have a compact pole at infinity. Similarly, it can be shown that if the admittance parameters have a compact pole at infinity, the residues of the impedance parameters at infinity will be compact. The denominator of the admittance parameters will have the same form as (13) if the highest degree of the numerator polynomials is n. The residues of the pole at infinity of the admittance parameters are given by (14) and (15) also holds. Thus, the residues at infinity of the impedance parameters given by (10) are also compact.

Let us assume that we are given a set of impedance parameters which has compact residues, a pole at the origin, and no common zeros. The impedance parameters can be expressed by (5) with $d_\circ = 0$; thus, the residues of the pole at the origin are

$$\overset{(\circ)}{k}_{11} = \frac{a_\circ}{d_1} \qquad \overset{(\circ)}{k}_{22} = \frac{b_\circ}{d_1} \qquad \overset{(\circ)}{k}_{12} = \frac{c_\circ}{d_1}\,. \tag{27}$$

Since the residues at the pole at the origin are compact, we have

$$\overset{(\circ)}{k}_{11}\,\overset{(\circ)}{k}_{22} - (\overset{(\circ)}{k}_{12})^2 = \frac{a_\circ b_\circ - c_\circ^2}{d_1^2} = 0\,. \tag{28}$$

The admittance parameters are expressed by (8) where e(s) has the form of (13) since the residues at infinity of the z parameters are compact. The residues at zero of the admittance parameters are

$$\overset{(\circ)}{\lambda}_{11} = \frac{b_\circ}{e_\circ} \qquad \overset{(\circ)}{\lambda}_{22} = \frac{a_\circ}{e_\circ} \qquad \overset{(\circ)}{\lambda}_{12} = \frac{c_\circ}{e_\circ} \tag{29}$$

and from (28) we have

$$\overset{(\circ)}{\lambda}_{11}\,\overset{(\circ)}{\lambda}_{22} - (\overset{(\circ)}{\lambda}_{12})^2 = \frac{b_\circ a_\circ - c_\circ^2}{e_\circ^2} = 0 \tag{30}$$

Thus, the residues at zero of the y parameters are compact.

Proof of Necessity:

Let us assume that we are given a set of impedance and admittance parameters that have compact residues and no common zeros. The residues of the impedance parameters at infinity cannot be zero; otherwise, there is a common zero at infinity. Thus, the impedance parameters can be expressed by (5) and it has been assumed that the residues at infinity are compact. It was shown previously that the admittance parameters will have a compact pole at infinity if the impedance parameters have compact residues at infinity. Thus, the admittance parameters must have a pole at infinity.

The residues of the admittance parameters at zero can be expressed by (29) and since the residues are compact (30) also holds. The polynomial d(s) is given by (9) with e(s) and d(s) interchanged. Since (30) holds, the polynomial $[n_{11}(s)n_{22}(s)-(n_{12}(s))^2]$ has no constant term and d(s) has a zero at $s = 0$. Thus, the impedance parameters have a pole at the origin. The numerator polynomials cannot have a zero at $s = 0$; otherwise, the parameters will have a common zero. The residues of the pole at the origin of the impedance parameters is given by (27); thus, (28) holds also since (30) is satisfied. This completes the proof of Theorem 2.

CONCLUSIONS

Necessary and sufficient conditions have been given for a set of impedance or admittance parameters of an RC two-port to satisfy either Slepian's or Dasher's definition of compactness when a compact set of admittance or impedance parameters is specified. The conditions can be applied by inspection of the zeros of the specified set of parameters.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to Mrs. Chris Kinstler and Mrs. Audrey Veschusio for their excellent secretarial assistance.

REFERENCES

1. N. Balabanian, Network Synthesis. New Jersey: Prentice Hall, 1958.
2. B. J. Dasher, "Synthesis of RC Transfer Functions as Unbalanced Two-Terminal Pair Networks," I.R.E. Transactions on Circuit Theory, CT-1, pp. 20-34, Dec. 1952.
3. E. A. Guillemin, Synthesis of Passive Networks. New York: John Wiley & Sons, 1957.
4. P. Slepian, "On the Compactness of an RC Quadripole," I.R.E. Transactions on Circuit Theory, CT-5, pp. 61-65, March, 1958.
5. M. E. VanValkenburg, Modern Network Synthesis. New York: John Wiley & Sons, 1960.
6. L. Weinberg, Network Analysis and Synthesis. New York: McGraw-Hill, 1962, p. 31.

LC TRANSFER FUNCTION SYNTHESIS BY STATE SPACE TECHNIQUES

TIEN HUA AND EDDIE R. FOWLER
Electrical Engineering Department, University of Missouri - Rolla
Rolla, Missouri 65401

Abstract - Previous work utilizing state space techniques in transfer function synthesis has only stated that these techniques can also be used for LC transfer function synthesis. This paper justifies that assertion since state space techniques are utilized to yield synthesis procedures for all physically realizable LC transfer functions, i.e. $Y_{12}(s)$, $Z_{12}(s)$, and $V_o(s)/V_{in}(s)$. The resulting two port networks are interconnected LC ladders with or without transformers. The synthesis procedure gives the ladder topology, equations for obtaining network component values, port locations, and transformer turns ratios. Also the procedure is systematized to allow programming on the digital computer. If desired it is possible to unify the component values in the interconnected ladder networks to facilitate integrated circuit fabrication of the synthesized transfer function.

STATE SPACE TECHNIQUE AND CHARACTERISTIC POLYNOMIAL

The state space technique is to obtain an A-matrix from the transfer function characteristic polynomial whose eigenvalues are the same as the roots of the characteristic polynomial. Consider an even or odd characteristic polynomial, D(s), of degree, $n = t + q$;

$$D(s) = s^t(s^q + b_{q-2}s^{q-2} + \ldots + b_2 s^2 + b_o) \tag{1}$$

Yarlagadda [1] has presented a technique that developes an A-matrix as shown whose eigenvalues are equal to the roots of D(s) and can also be presented algebraically as

$$\frac{d}{dt}\left[\begin{array}{c} I'_{c\ell} \\ \hline V'_{bc} \end{array}\right] = \left[\begin{array}{c|c} 0 & K_{12} \\ \hline -K_{12}^T & 0 \end{array}\right]\left[\begin{array}{c} I'_{c\ell} \\ \hline V'_{bc} \end{array}\right] = \left[\begin{array}{c|c} -L_c^{-\frac{1}{2}}B_{23}R_b B_{23}^T L_c^{-\frac{1}{2}} & -L_c^{-\frac{1}{2}}B_{22}C_b^{-\frac{1}{2}} \\ \hline C_b^{-\frac{1}{2}}B_{22}^T L_c^{-\frac{1}{2}} & -C_b^{-\frac{1}{2}}B_{12}^T G_c B_{12} C_b^{-\frac{1}{2}} \end{array}\right]\left[\begin{array}{c} I'_{c\ell} \\ \hline V'_{bc} \end{array}\right] \tag{2}$$

where K_{12} is a submatrix of order (mxr) and $n = m + r$ and whose only nonzero k_i-entries are located at (1,1), (2,1), (2,2), (3,2), (3,3), ..., (m,r).

Equating like submatrices of the A-matrices of (2) yields a set of matrix equations that 1) can be decomposed [2] to give the entries of the B_{ij} matrices, and that 2) result in a set of (n) nonlinear algebraic equations with $(n+1)$ unknowns. The B_{ij} matrices yield the source free ladder networks of Figure 1 whose free modes are equivalent to the roots of D(s) in (1). The set of nonlinear equations have a possible solution, assuming $L_1 = 1$ henry, of:

$$L_i = (k_1 k_3 \ldots k_{2i-3})^2/(k_2 k_4 \ldots k_{2(i-1)})^2 \;;\; 1 < i \leq m \tag{3a}$$

$$C_i = (k_2 k_4 \ldots k_{2(i-1)})^2/(k_1 k_3 \ldots k_{2i-1})^2 \;;\; 1 \leq i \leq r \tag{3b}$$

SHORT CIRCUIT TRANSFER ADMITTANCE, $Y_{12}(s)$

The discussion above has presented the procedure by which the characteristic polynomial of the transfer function is synthesized. Now the numerator polynomial must be considered. Only the short circuit transfer admittance, $Y_{12}(s)$, will be considered as the synthesis procedures for it are very similar to those for $Z_{12}(s)$ and $V_o(s)/V_{in}(s)$. Also since there are only two types of LC transfer functions, these will be identified as Case II $Y_{12}(s)$ and Case III $Y_{12}(s)$.

Case II $Y_{12}(s)$. This short circuit transfer admittance has an even polynomial in the numerator and an odd characteristic polynomial. Therefore x is even and n is odd and $Y_{12}(s)$ can be written as

$$Y_{12}(s) = (a_x s^x + a_{x-2} s^{x-2} + \ldots + a_2 s^2 + a_o)/(s^n + b_{n-2} s^{n-2} + \ldots + b_3 s^3 + b_1 s) \quad (4)$$

$$i = (n-x-1)/2 \quad (5)$$

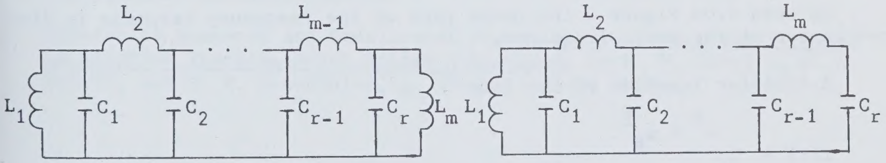
To obtain a Case II $Y_{12}(s)$, as shown in (4), with the correct numerator degree, x, insert in Figure 1(a) a voltage source in series with inductor, L_1, and insert another in the bottom "line" of the C_i, L_{i+1}, and C_{i+1} mesh where (i) is calculated as shown in (5). To obtain the a_x coefficient given in (4), use the L_i and C_i values determined in (3) to obtain a $Y_{12}(s)$ and a_x' coefficient for the two port ladder network just obtained above. Now magnitude scale the L_i's and C_i's to obtain the desired a_x coefficient.
Case III $Y_{12}(s)$. This short circuit transfer admittance has an odd polynomial in the numerator and an even characteristic polynomial. Therefore in (4) x is odd and n is even. To obtain a Case III $Y_{12}(s)$ with the correct numerator degree, x, follow the same procedure as given for the Case II $Y_{12}(s)$ except use Figure 1(b).

The procedure by which the rest of the a_i coefficients in the numerator of $Y_{12}(s)$ are satisfied has previously [3] been presented in great detail. Briefly this procedure is to parallel (x) ladder networks, of the types shown in Figure 1, with the sources added in such a way that when each ladder $Y_{12}^{(i)}(s)$, $1 \leq i \leq x$, is added together the resultant $Y_{12}(s)$ will have the same a_i and b_i coefficients as the original $Y_{12}(s)$. When paralleling the (x) ladder networks, it may be necessary to add transformers to obtain positive a_i coefficients or to satisfy the validity test. A transformer could be added to each ladder network to provide the magnitude scaling. This results in each ladder network having unified component value, i.e. the L_1 component value of each ladder will be the same, the C_1 component value of each ladder will be the same, etc.

Open circuit transfer impedance, $Z_{12}(s)$, and voltage transfer functions, $V_o(s)/V_{in}(s)$, require synthesis procedures very similar to those described above. The voltage transfer functions are the only ones that definitely require a transformer for every ladder network. All of these synthesis procedures are readily programmable in Fortran IV. A much more detailed paper presenting this material is available from the authors.

We wish to acknowledge Mr. Ken Litteken who did much of the first research work on this topic. Also the National Science Foundation is acknowledged for providing the financial support for this research under NSF Grant GK-5466.

[1] Yarlagadda, R. "An Application of Tridiagonal Matrices to Network Synthesis." SIAM Journal of Applied Mathematics, (November, 1968), 1146.
[2] Cederbaum, I. "Applications of Matrix Algebra to Network Theory." IRE Transactions on Circuit Theory, vol. CT-6, (May, 1959), 127-137.
[3] Fowler, E. R. "A Unified State-Space Approach to RLCT Two-Port Transfer Function Synthesis." (unpub. Ph. D. thesis, Oklahoma State University, 1969).



(a) n is odd n = m + r (b) n is even

Figure 1. Source Free Ladder Networks

SOME CHARACTERISTICS OF A COMMUTATED NETWORK FILTER

T. R. NILSEN
Royal Norwegian Navy

G. J. THALER
Naval Postgraduate School
Monterey, California

A commutated network is an electrical network, either active or passive, consisting of resistors and reactive elements which are switched or commutated in a periodic manner. In most practical cases the reactive elements are capacitors. There are several configurations of a commutated network used in the field of electrical engineering. The configuration used here is that of commutated capacitors used in the feed-back path around an operation amplifier. Thus by applying commutation action to the capacitors in a lag network, the low-pass frequency response of that network can be transformed into that of a highly selective bandpass filter. The center frequency of the bandpass characteristic will be the same as the commutation frequency.

The objective of the RC-commutated network is to construct a signal of the same frequency but of opposite phase as the input signal. If this is the case the sum of the input signal and the output signal has zero component of the input frequency. Figure 1 shows how this can be implemented.

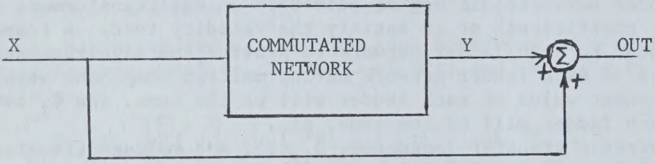


Figure 1. Filter Configuration

FREQUENCY RESPONSE OF FILTER

An investigation of the phase difference between output and input signals was done for frequencies within one decade on either side of the notch frequency, ω_n.

Figure 2 shows the plot of phase versus frequency as obtained. The phase starts at zero degrees for low frequencies and goes to -90 degrees for frequencies slightly less than ω_n, jumps to +90 degrees for frequencies slightly above ω_n, and goes back to zero degrees for high frequencies.

APPROXIMATION OF FILTER TRANSFER FUNCTION

As seen from Figure 2 the phase part of the frequency response is discontinuous at the notch-frequency.

A transfer function of the form

$$T(s) = \frac{s^2 + \omega_n^2}{s^2 + 2\zeta \omega_n s + \omega_n^2}$$

is seen to have the same phase characteristic. By varying the value of the damping factor the curvature of the phase curve can be changed.

Figure 2 shows a comparison of the phase curves of T(jω) with the phase of the commutated network. It is apparent that ζ can be chosen so that the phase of T(jω) closely approximates that of the commutated filter. A numerical correlation has not been attempted here. The phase obtained with the commutating network depends on the value of the network time constant, and further study is needed.

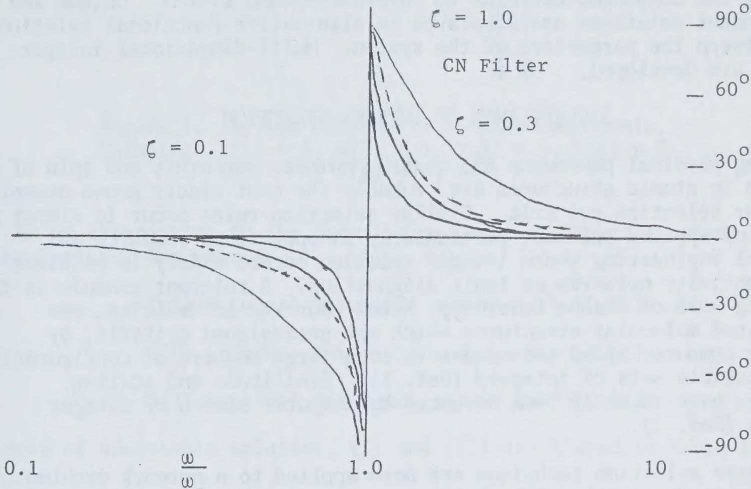


Figure 2. Bode Plot of T(s) phase and C.N. phase

APPLICATIONS OF FILTER IN SERVOSYSTEMS

Computer simulations were done with the filter used in conjunction with servosystems. The objective of the filter was to suppress the component of frequency ω_n in the output signal, where ω_n was the frequency of a lightly damped resonance in the compensated system. Design calculations used the approximate transfer function, T(s), and simulation results indicated that T(s) is an adequate model for design purposes.

LIST OF REFERENCES

1. Franks, L. E. and Sandberg, I. W., "An Alternative Approach to the Realization of Network Transfer Functions: The N-Path Filter", The Bell System Technical Journal, p. 1321-1324, September 1960.

2. Astrionics Research and Development Report No. 3, Recent Developments in Adaptive Tracking Notch Filter Techniques, by M. T. Borelli, S. N. Caroll, and H. H. Hosenthien, p. 39-40, 1 March 1965.

DISPLAY OF FUNCTIONAL RELATIONSHIPS BETWEEN INTEGER SELECTION CRITERIA

William W. HAPP
U. S. Army Corps of Engineers, Champaign, Illinois and
University of Illinois, Urbana, Illinois

ABSTRACT

Selection rules are established for the number of terminals in combining two component networks to form a compound system. Unique and non-redundant solutions are displayed in alternative functional relationships between the parameters of the system. Multi-dimensional integer displays are developed.

SIGNIFICANCE OF INTEGER SELECTION

Among physical phenomena the quantum numbers for orbit and spin of electrons in atomic structures are probably the most widely known examples of integer selection criteria. Similar selection rules occur in almost all branches of applied science, particularly in operations research and industrial engineering where integer solution occurs widely in problems such as activity networks or fault diagnostics. A relevant example is the pioneering work of Jushua Lederberg, Nobel Laureate in Medicine, who investigated molecular structures which met preassigned criteria, by utilizing computer-aided techniques to scan large numbers of configurations for permissible sets of integers (Ref. 1). Similitude and scaling techniques have recently been advanced by computer search of integer solutions (Ref. 2).

Integer selection technique are here applied to a network problem: Two "component" networks with at least two terminals are combined to form a resultant "system" network with at least two terminals, such that at least one common terminal is shared by each component. Desired is a display between the significant parameters of the system.

ALTERNATIVE PROBLEM STATEMENTS

All variable to be defined are integers. Assume k sets, such that

$X(k)$ is a series of successive non-negative integers

$Y(k)$ is the maximum of the variable $X(k)$

$Z(k)$ is the minimum of the variable $X(k)$ and

$R(k)$ is the range of values of the variable $X(k)$

It follows that $R(k) = Y(k) - Z(k) + 1$.

The above network problem is described by:

$X(1)$ terminals external to the system

$X(2)$ terminals of one component

$X(3)$ terminals of the other component

$X(4)$ internal connections common to $X(2)$ and $X(3)$

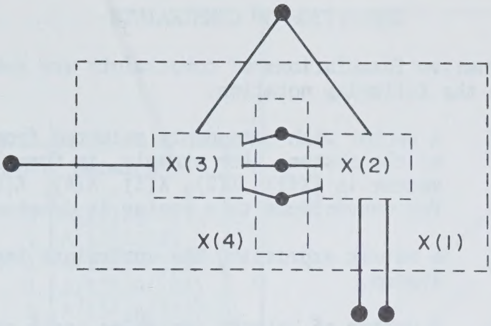


Figure 1. System with Multiterminal Components.
Example: X(1) = 4, X(2) = 6, X(3) = 5, X(4) = 3
X(5) = 1, X(6) = 3, X(7) = 2

As an alternative to the formulation given in Figure 1, other variables can be defined, for example:

X(5) external terminals common to both X(2) and X(3)

X(6) external terminals common only to X(2)

X(7) external terminals common only to X(3)

Two sets of admissible solution, (I) and (II) are listed in Table 1.

Input (I)	X(1)	222222	333333	44444
	X(2)	233444	233444	33444
	X(3)	223234	223234	23234
Output (I)	Z(4)	011222	000111	00000
	R(4)	212112	122122	12223
Input (II)	X(2)	22	33333	4444444444
	X(3)	22	22333	2223334444
	X(4)	01	01012	0120120123
Output (II)	Z(1)	22	32322	4324324322
	R(1)	21	22331	2213314431

Table 1. Alternative Admissible Solutions.

Input (I) constitutes a possible set of independent variables for which Output (I) is the set of corresponding dependent variables. Some formulations provide significantly better insight than others. A large variety of input/output formulations are possible.

DEFINITION OF CONSTRAINTS

Three alternative formulations of constraints are summarized in Table 2, based on the following notation:

$\underline{X}$	a vector with components selected from the variables of the system. For example, in formulation (I) this vector is [X(1), X(2), X(3), X(4), X(0)] with X(0)=1. For convenience this vector is denoted as X(1,2,3,4,0).
$\underline{C}$	a matrix expressing the contraints imposed upon the system.
$\underline{S}$	a vector of "slack" variables, each component consisting of non-negative integers S = 0,1,2,...

This notation describes the constraints imposed upon the system as

$$\underline{X}\ \underline{C} = \underline{S}$$

For example, in formulation (I) the first two rows of $\underline{C}$ read:

$$X(1) \geq 2 \text{ and } X(2) \geq 2$$

Each of the formulations (I), (II) and (III) denotes seven inequalities relating four variables and provides a necessary and sufficient problem statement. The equivalence of (I), (II) and (III) can be demonstrated by transformations of variables.

	Formulation(I)	Formulation(II)	Formulation(III)
$\underline{X}$	[X(1, 2, 3, 4, 0)]	[X(5, 2, 3, 4, 0)]	[X(5, 6, 7, 4, 0)]
$\underline{C}$	$\begin{bmatrix} 1 & . & . & . & -2 \\ . & 1 & . & . & -2 \\ . & . & 1 & . & -2 \\ . & . & . & 1 & . \\ . & 1 & -1 & . & . \\ 1 & -1 & . & 1 & . \\ 1 & -1 & -1 & 2 & . \end{bmatrix}$	$\begin{bmatrix} 1 & . & . & . & . \\ . & 1 & . & . & -2 \\ . & . & 1 & . & -2 \\ . & . & . & 1 & . \\ . & 1 & -1 & . & . \\ -1 & . & 1 & . & . \\ 1 & . & . & 1 & -1 \end{bmatrix}$	$\begin{bmatrix} 1 & . & . & . & . \\ . & 1 & . & . & -1 \\ . & . & 1 & . & . \\ . & . & . & 1 & . \\ 1 & 1 & 1 & . & -2 \\ . & 1 & -1 & . & . \\ 1 & . & . & 1 & -1 \end{bmatrix}$

Table 2. Alternative Formulations of Constraints.

MULTI-DIMENSIONAL DISPLAY

Under the constraints $\underline{C}$ each set of values $\underline{X}$, for example X(1), X(2), X(3), X(4), admits only two possible answers: consistent or inconsistent. A "truth table" for X can be established and contains admissible values as displayed in Table 3. A more economical display of admissible solutions results, if the auxiliary variables Y(k), Z(k) and R(k) are introduced.

X(1)	X(2)	X(3)	X(4)	Network
2	2	2	0	A(12)B(12)
			1	A(13)B(23)
2	3	2	1	A(123)B(23)
2	3	3	1	A(123)B(123)
			2	A(134)B(234)
2	4	2	2	A(1234)B(34)
2	4	3	2	A(1234)B(134)
2	4	4	2	A(1234)B(1234)
			3	A(1345)B(2345)

Table 3. Enumeration of Admissible Network Configurations.

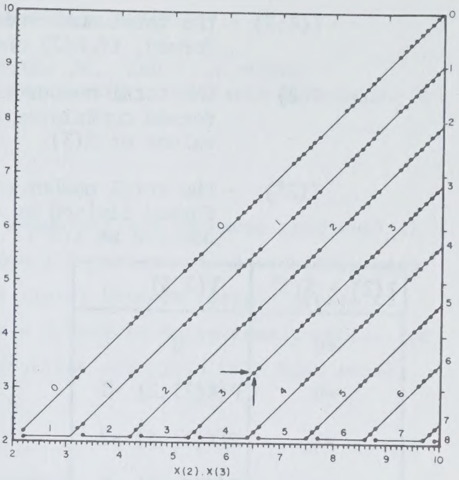


Figure 2. Four-Dimensional Display Example: X(2) = 6, X(3) = 5, R(1) = 3, Z(1) = 3 Running Parameter X(4) = 3

Since the solutions are a set of small integers, two or more variables can be displayed simultaneously along the same axis. Figure 2 illustrates how several hundred admissible solutions from Table 1 can be presented in a single four-dimensional display. The display in Fig. 2 plots admissible values of Z(1).R(1) as a function of X(2).X(3), with X(4) as a running parameter.

Certain combinations of variables often displays functional relationships between selection criteria particularly effectively. In turn an effective display of selection criteria frequently permits additional insight into the problem, thereby leading to a preferred set of variables.

ALGORITHMS FOR COMBINING OF MULTITERMINAL NETWORKS

To formulate strategies for combining or "tearing" multiterminal networks, a listing of non-redundant combinations of X(2) and X(3) to form X(1) is needed. Table 3 lists these permissible combinations for X(1) = 2. For example A(134)B(234) denotes the two-terminal system C(12) with two internal connections, nodes 3 and 4, which couples the three-terminal component A to the three-terminal component B. The value of the digit in A and B refers to terminal position in C, while position of the digit in A and B denotes the terminal position in A and B.

To establish algorithms and strategies for analyzing large systems in terms of their constituent subsystems, it is essential to analyze integer selection criteria.

APPLICATIONS TO NETWORK COMBINATORICS

Integer selection rules are essential to perform summations over specific classes of objects (Ref. 3). Three sample summations will illustrate this approach using multiterminal networks as an example:

T(2,3) - the total number of network combinations which can be formed, if X(2) and X(3) are specified.

T(2) - the total number of network combinations which can be formed containing a specified X(2) and all admissible values of X(3).

T(2*) - the total number of network combinations which can be formed limited by a maximum specified value of X(2) denoted as X(2*).

X(2) - X(3)	T(2,3)
<0	0
=0	P(X(2),2) -3
=1	P(X(2),2) -2
≥2	P(X(2),2) -1

Table 4. Values of T(2,3).

X(2) or X(2*)	2	3	4	5	6 . . .
T(2)	(3)	11	25	45	72 . . .
T(2*)	3	14	39	84	156 . . .

Table 5. Unique Networks.

Defining the Poisson function $P(J,K) = (J + K)!/(J!K!)$, Table 4 summarizes values for T(2,3) obtained from Table 2 by inspection. By definition T(2) is the summation of T(2,3) from X(3) = 2 to X(3) = X(2) obtainable from summation formulas for Poisson functions:

$$T(2) = P(X(2),3) - X(2) + 3!$$

Similarly T(2*) is the summation of T(2) from X(2) = 2 to X(2) = X(2*).

$$T(2^*) = P(X(2^*),4) - P(X(2^*) + 5,2) + 4!$$

These results can be verified for small values listed in Table 5, which were obtained by enumeration.

The formula for T(2) gives an erroneous result for X(2) = 2, namely T(2) = 2. The correct result is obtained by enumeration as T(2) = 3.

REFERENCES

(1) J. Lederberg "*System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs.*" Interim report to NASA Grant Nc 81-60, Part II, December 1965.

(2) W. W. Happ, *Computer-Oriented Procedures for Dimensional Analysis*, Journal of Applied Physics, Vol. 38, No. 10, pp 3918-3927, Sept. 1967.

(3) W. W. Happ, *Combinatorial Analysis of Multi-Terminal Devices*, IEEE Transactions, Vol. SSC-3, No. 1, pp 21-27, June 1967.

A RELATIONAL DATA STRUCTURE FOR DOCUMENT RETRIEVAL*

Y. H. CHANG, R.T. CHIEN, N.M. ESSER, JR.†, AND F. A. STAHL
Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois

Abstract

In this paper a relational data structure for document retrieval is presented. The structure is based upon a hierarchic tree structure with superimposed linkages (called lateral links) between various groups of nodes in the tree. These linkages have proven to be extremely useful for interactive information retrieval operations and yield both high recall and high precision rates.

I. INTRODUCTION

In recent years, a great deal of work has been done in the area of document retrieval to improve their effectiveness in such aspects as: indexing, file organization, searching strategy, clustering, interactive techniques, etc. Among these areas interactive techniques have been found to be one of the most effective ways to improve the overall system performance.

In this paper a relational data structure for document retrieval is presented. The structure is based upon a hierarchic tree structure with superimposed linkages called lateral links between various groups of nodes in the tree. These linkages are useful for retrieval operations that yield both high recall and high precision rates.

In order to show the effectiveness of this approach a relational data structure that was developed and implemented for documents dealing with the subject of Coding Theory is presented. The implementation is in ISL (Information Search Language), an interactive language for information retrieval [7,10].

Comments on a number of interactive document retrieval systems are presented in section II. In section III the relational data structure is described along with examples of its effective use.

---

[*] Supported by the Office of Education under Contract No. OE C-1-7-071213-4557 and by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract DAAB-07-67-C-0199.

[†] Presently at National Cash Register, Dayton, Ohio

II. COMMENTS ON SOME INTERACTIVE DOCUMENT RETRIEVAL SYSTEMS

Experiments on the SMART system by Salton [9,14,15] have clearly shown that interactive techniques are the most effective for improving system performance. Salton adopts the interface of the user after the initial search operation has been completed. The evaluation of the relevance of the retrieved documents by the user is used only to improve the result of additional searches. The extra search operations could be avoided if the interaction of the user could be applied beforehand.

Lancaster's report [13] of the evaluation result on MEDLARS indicates that most of the source of irrelevant recall is from a poorly stated request, and Atherton, et al.[1] also noted that the query formulation is the biggest source of retrieval failure.

The BOLD system described by Burnaugh [3,4] shows subject categories to the user first, and then displays the titles, authors, and abstracts of the documents in any selected category on the CRT screen, or displays the relation matrix of document references and index terms for the user to select the relevant documents. The AESOP system [2] uses hierarchic tree structures to handle the category titles, and tables to handle the items in each category. Both the BOLD and AESOP systems have the same approach and are different only in the form of presentation. Neither of them helps the user more than restricting him in a machine provided guide. Furthermore, if the number of documents in one category is too large, the system becomes ineffective.

Kellogg's CONVERSE system [11,12] and Cautin's New English system [5] both remove the noise words in the query, and question the user about the words which they could not recognize instead of trying to help the user to make a more valid or accurate query.

III. THE RELATIONAL DATA STRUCTURE

The tree structure is the underlying structure upon which the relational structure is based. The reasoning behind this approach is that hierarchic categorization is straightforward and can represent much of the information that is needed for effective information retrieval. In addition, the user of a system based on this categorization scheme can familiarize himself with this part of the structure fairly easy. However, some information just cannot be represented using such a neat and simple structure.

In order to overcome some of the inadequacies inherent in the hierarchic tree structure, one must add to it additional information that *relates* certain nodes of the tree structure under the proper conditions to certain other nodes. Such relations are called *lateral links* [8] because they generally relate information *laterally* throughout the tree structure from one group of nodes to another. The lateral links give the structure greater flexibility for use in processing requests. They can be used to select new, possibly relevant areas to include in a search, as well as to discriminate between related, but irrelevant topics.

Consider the following examples with respect to the tree structure given in fig. 1:

1. If the nodes *optimum codes* and *high rate* are given, then through the use of lateral links *meet upper bound* and *Hamming bound* should be suggested.
2. If the nodes *binary* and *error evaluator* are given, then the node *error locator* should be suggested.

With the more specific information obtained through the use of the lateral links such as these more intelligent responses are possible. This yields more reasonable approach to the problem.

The lateral links can be defined symbolically as follows:

$$f(N) \longrightarrow g(M)$$

where N and M are sets of nodes, and f and g are logical operations on nodes. The interpretation, here, is that if a certain condition f on the nodes N holds true, then a suggestion of the form g over the nodes M is put forth by the system.

Thus, in example 1. above, if f_1(*optimum codes*, *high rate*) holds true where $f_1(a,b) = a \wedge b$, then *meet upper bound* and *Hamming bound* are suggested.

It is not reasonable to assume that the query will contain as key-words the descriptors necessary for efficient retrieval. For example, suppose the user wants 'papers dealing with *codes* which meet the *Hamming bound*.' Strictly by using the key-words, he may get all the documents which are concerned with *codes* and *Hamming bound*. But, through the lateral links, he gets only *Perfect codes* and *Optimum codes of high information rate* if desired.

Documents dealing with *perfect codes* and *optimum codes* may be included in the group of retrieved documents which talk about *codes*, only if all documents dealing with *codes* are retrieved. This query would, hence, not be specific enough for a key-word based retrieval system. But, on the other hand, it would be quite specific with the use of lateral links.

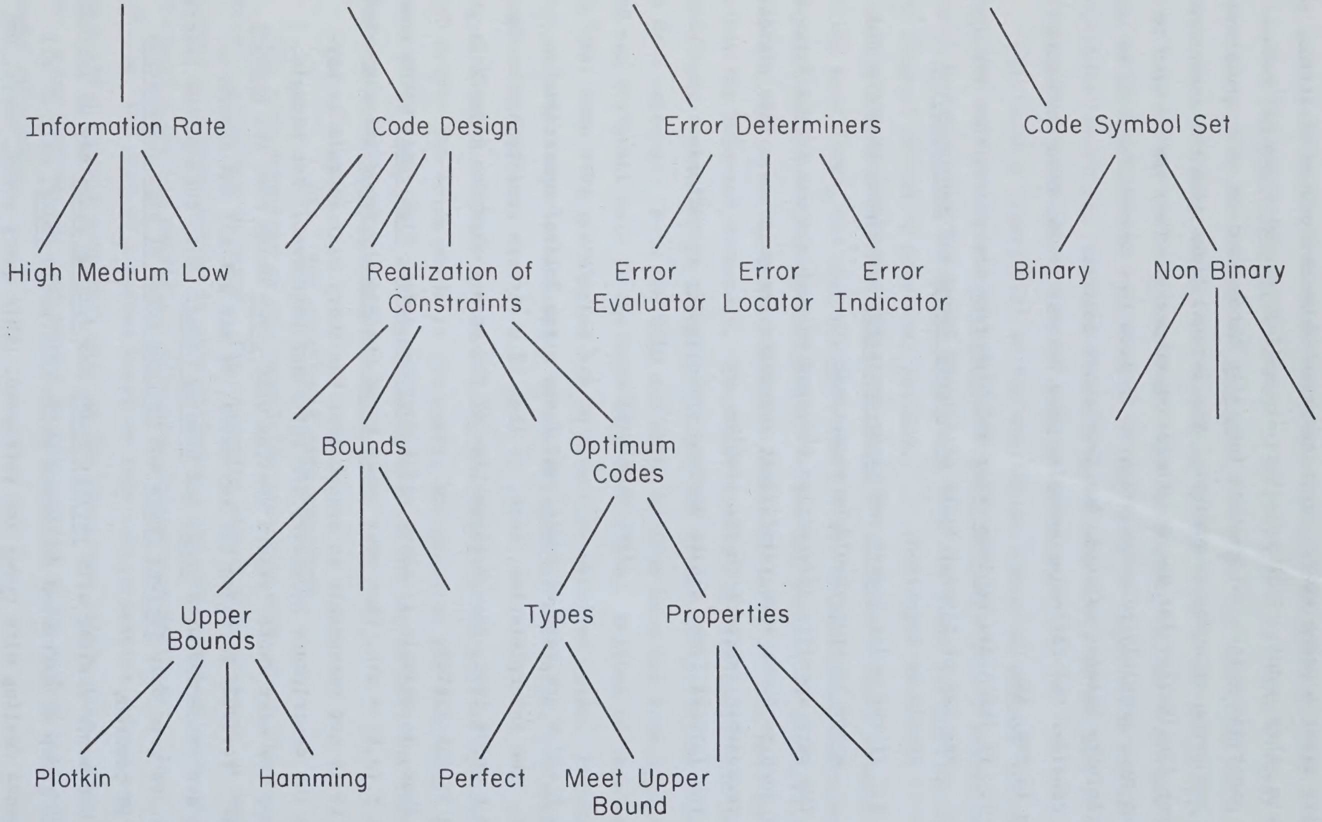

Information Rate
High
Medium
Low
Code Design
Realization of Constraints
Bounds
Upper Bounds
Plotkin
Hamming
Optimum Codes
Types
Perfect
Meet Upper Bound
Properties
Error Determiners
Error Evaluator
Error Locator
Error Indicator
Code Symbol Set
Binary
Non Binary

Figure 1 A Partial Tree for Coding Theory

Moreover, in this example, codes might not have been accepted at all as a key-word since in Coding Theory most documents deal with codes. In that case, no documents dealing with codes could have been retrieved.

In order to perform experiments with a relational data structure for retrieval purposes an interactive system has been implemented [6]. In this implementation a variety of interesting features are available for the user after the specification of the initial query.

The user may add any new terms that he feels will improve the query, or he may delete any terms that he has decided do not really reflect his interest (especially if a lateral link suggests it, and he agrees). This is all accomplished via light-pen control or typewriter at a display console.

For any term the user may also inquire about its position in the tree structure. That is, he may ask about the nodes immediately above or below the node in question. Of course, throughout this interactive portion of the retrieval process the lateral links that are available to the system are suggesting the addition of new nodes or the deletion of others. Once the user has decided that the altered query reflects what he really wants he can get a hard copy of the query and initiate the actual retrieval operation.

IV. CONCLUSION

Using the relational data structure to help the user formulate a query has proved to be an effective means of attaining higher precision and recall rates through an intelligent dialogue prior to the initial search. The lateral links supply information that can make the query more precise, and thus save the user and the retrieval system much valuable time.

The use of the relational data structure for information retrieval does not prohibit the use of other techniques like Salton's that do analysis after the documents have been retrieved. The use of this type of aid in the retrieval process can only improve the overall quality of the system.

REFERENCES

1. Atherton, P., King, D. W., Freeman, R. R., "Evaluation of the Retrieval of Nuclear Science Document References Using the UDC as the Indexing Language for a Computer-Based System," Report AIP/UDC-8, American Institute of Physics, New York, May, 1968.

2. Bennet, E., Haines, E., Summer, J., "AESOP: A Prototype for On-Line User Control of Organizational Data Storage, Retrieval, and Processing," AFIPS Conf. Proc. - 1965 Fall Joint Computer Conf., Vol. 27, Part 1, Spartan Books, Washington, D.C., 1965, pp. 435-456.

3. Burnaugh, H. P., "Introduction to the BOLD Retrieval System," Report TM-2306/002/03, Systems Development Corp., Santa Monica, Calif., Jan., 1966.

4. Burnaugh, H. P., "Retrieval Program for the BOLD System," Report TM-2306/002/00, Systems Development Corp., Santa Monica, Calif., Jan., 1966.

5. Cautin, H., Lowe, T. C., Rapp, F., Rubinoff, M., "An Experimental On-Line Information Retrieval System," Moore School of Electrical Engineering, University of Penna., Phila., Penna., April, 1967.

6. Chang, Y. H., "The Implementation of a Relational Document Retrieval System," Coordinated Science Lab., University of Illinois at Urbana-Champaign, Urbana, Illinois, Oct., 1970.

7. Chien, R. T., Ray, S. R., Stahl, F. A., "ISL - A New Information Retrieval Language," in Proceedings of the Third International Symposium of Computer and Information Science, edited by Julius Tou, Academic Press, New York, 1970 (tp appear).

8. Esser, N. M., Jr., "A Relational Structure for Document Retrieval in Coding Theory," Coordinated Science Lab., University of Illinois at Urbana-Champaign, Urbana, Illinois, May, 1970.

9. Ide, E., Salton, G., "Interactive Search Strategies and Dynamic File Organization in Information Retrieval," Report ISR-16, Dept. of Computer Science, Cornell University, Ithica, New York, Sept., 1969.

10. Kelley, K. C., Ray, S. R., Stahl, F. A., "ISL - A String Manipulating Language," Report R-407, Coordinated Science Lab., University of Illinois at Urbana-Champaign, Urbana, Illinois, Feb., 1969.

11. Kellogg, C. H., "On-Line Translation of Natural Language Questions into Artificial Language Queries," Systems Development Corp., Santa Monica, Calif., April, 1967.

12. Kellogg, C. H., "CONVERSE - A System for the On-Line Description and Retrieval of Structured Data Using Natural Language," Systems Development Corp., Santa Monica, Calif., May, 1967.

13. Lancaster, F. W., "Evaluation of the MEDLARS Demand Search Service," National Library of Medicine, Bethesda, Md., Jan., 1968.

14. Lesk, M. E., Salton, G., "Interactive Search and Retrieval Methods Using Automatic Information Displays," Report ISR-14, Dept. of Computer Science, Cornell University, Ithica, New York, Oct., 1968.

15. Salton, G., Automatic Information Organization and Retrieval, McGraw-Hill, New York, 1968.

OPTIMAL DELAYED DECISION PROCEDURES FOR UNSUPERVISED PATTERN RECOGINITION SYSTEMS

A. G. JAFFER AND S. C. GUPTA
Information and Control Sciences Center
SMU Institute of Technology
Dallas, Texas 75222

INTRODUCTION

One of the central problems in pattern recognition systems is to obtain optimal decision rules for sequentially classifying an observation vector X_k as belonging to one of M categories, based on the previous observation sequence $X_1 \ldots X_{k-1}$ of unknown true classifications. Due to inadequate prior knowledge of the problem structure, the system is required to progressively adapt its structure to improve classification, in accordance with its improved knowledge of the problem structure. Such problems have been fairly well treated in the literature (1-5). The usual formulation of the problem, however, requires that a decision be made on an observation vector X_k immediately as X_k is received. However, deferring a decision on X_k to a stage $k+\Delta$, where Δ is a fixed delay, can result in a significant reduction of the Bayes risk because of: 1) The sequential dependence between hypotheses and 2) The greater knowledge of the internal structure available after data up to a stage $k+\Delta$ has been observed rather than only up to k.

The purpose of this paper is to obtain the Bayes optimal delayed decision rules for unsupervised recognition systems with dependent hypotheses. While it may be expected that the form of the Bayes rule is modified in an obvious manner to account for the delay in the classification of X_k, however the cardinal purpose of this work is to obtain the decision rules in a recursive, finite memory form. The introduction of a fixed delay Δ in the processing of X_k, adds significantly to the task of obtaining recursive, non-growing optimal solutions. A parametric model is considered which is broad enough to include context dependence between successive hypotheses.

FORMULATION OF THE CLASSICATION PROBLEM

Let $\underline{X}_k = (X_1, X_2, \ldots, X_k)$ be a sequence of unclassified observation vectors and let $\underline{\omega}_k = (\omega_1, \omega_2, \ldots, \omega_k)$ be the corresponding sequence of hypotheses. Each ω_k is a random variable whose sample space Λ consists of M elements corresponding to the M possible categories.

Two cases are considered in this paper: Case 1. It is assumed that the probability density function $p(X_k|\omega_k^j)$ is completely known for each k and j = 1, 2,...,M. The observations, conditioned on the hypotheses, are assumed to be independent. Specifically,

$$p(\underline{X}_k|\underline{\omega}_k) = \prod_{i=1}^{k} p(X_i|\omega_i) \tag{1}$$

To allow for context dependence, the hypotheses are assumed to form a first order Markov chain with known transition probabilities, i.e.,

$$P(\omega_k|\underline{\omega}_{k-1}) = P(\omega_k|\omega_{k-1}) \tag{2}$$

Case 2. The functional form of the p.d.f. of X_k conditioned on ω_k is assumed to be known with a set of unknown constant parameters θ; the observations conditioned on ω and θ are assumed to be independent, i.e.,

$$p(\underline{X}_k|\underline{\omega}_k,\theta) = \prod_{i=1}^{k} p(X_i|\omega_i,\theta) \tag{3}$$

Context dependence is accounted for by modeling the sequence of hypotheses as a first order Markov chain as in Case 1.

OPTIMAL DELAYED DECISION POLICIES

If $C(i|j) \geq 0$ is the cost of making decision i when j is true where

$C(i|j) \geq C(j|j)$, the quantity to be minimized for each stage k is the Bayes risk

$$R_k = \int_{S^{k+\Delta}} \sum_{\underline{\omega}_{k+\Delta} \in \Lambda^{k+\Delta}} C(i|j)\, p(\underline{X}_{k+\Delta}|\underline{\omega}_{k+\Delta})\, P(\underline{\omega}_{k+\Delta})\, d\underline{X}_{k+\Delta} \qquad (4)$$

where Δ is the delay (fixed), S is the sample space of the observation vector X, $S^{k+\Delta}$, $\Lambda^{k+\Delta}$ are the $(k+\Delta)$-fold Cartesian product of S and Λ respectively and $P(\underline{\omega}_{k+\Delta})$ is an a priori distribution over $\Lambda^{k+\Delta}$. Hence a risk minimizing decision rule consists of choosing that hypotheses i which minimizes

$$Q(i) = \sum_{j=1}^{M} C(i|j)\, P(\omega_k^j|\underline{X}_{k+\Delta})\, p(\underline{X}_{k+\Delta}) \qquad (5)$$

Thus the optimal adaptive system computes $P(\omega_k^j|\underline{X}_{k+\Delta})$, $j = 1,2,\ldots,M$ and chooses category i so as to minimize Q(i).

RECURSIVE SOLUTIONS

Case 1. We can write the expression for Q as

$$Q = \sum_{\omega_k} C(i|j)\, p(\underline{X}_{k+\Delta},\omega_k)$$

$$p(\underline{X}_{k+\Delta},\omega_k) = \Big\{ \sum_{\omega_{k+1}} \cdots \sum_{\omega_{k+\Delta}} p(X_{k+1}\cdots X_{k+\Delta}|\omega_{k+1}\cdots\omega_{k+\Delta})\, P(\omega_{k+1}\cdots\omega_{k+\Delta}|\omega_k,\underline{X}_k) \Big\}\, p(\underline{X}_k,\omega_k) \qquad (6)$$

Now $p(X_{k+1}\cdots X_{k+\Delta}|\omega_{k+1}\cdots\omega_{k+\Delta}) = \prod_{i=1}^{\Delta} p(X_{k+i}|\omega_{k+i})$ (7)

is a known quantity and

$$P(\omega_{k+1}\cdots\omega_{k+\Delta}|\omega_k,\underline{X}_k) = P(\omega_{k+\Delta}|\omega_{k+\Delta-1}) \cdot\cdot\cdot P(\omega_{k+1}|\omega_k) \qquad (8)$$

is known from the a priori knowledge of the transition probability law of the Markov chain. It remains to obtain a recursive relation for $p(\underline{X}_k,\omega_k)$.

$$p(\underline{X}_k,\omega_k) = p(X_k|\omega_k) \sum_{\omega_{k-1}} P(\omega_k|\omega_{k-1})\, p(\underline{X}_{k-1},\omega_{k-1}) \qquad (9)$$

Thus the computation for $p(\underline{X}_k,\omega_k)$ is recursive and starts initially with $p(X_1|\omega_1)P(\omega_1)$ where $P(\omega_1)$ is known a priori. $p(\underline{X}_{k+\Delta},\omega_k)$ is then obtained from (6), (7), and (8). The resulting optimal system is of fixed memory.

Case 2. We expand $P(\omega_k|\underline{X}_{k+\Delta})$ as a mixture in the unknown parameter θ where θ takes on values $\theta^1,\theta^2,\ldots\theta^R$

$$P(\omega_k|\underline{X}_{k+\Delta}) = \sum_{j=1}^{R} P(\omega_k|\underline{X}_{k+\Delta},\theta^j)\, P(\theta^j|\underline{X}_{k+\Delta}) \qquad (10)$$

thus the optimal adaptive system is characterized by the evolution of the conditional probabilities $P(\omega_k|\underline{X}_{k+\Delta},\theta)$ and $P(\theta|\underline{X}_{k+\Delta})$

$$P(\omega_k|\underline{X}_{k+\Delta},\theta) = p(X_{k+1}\cdots X_{k+\Delta}|\omega_k,\theta,\underline{X}_k) \cdot p(\omega_k|\underline{X}_k,\theta)/p(X_{k+1}\cdots X_{k+\Delta}|\theta,\underline{X}_k) \qquad (11)$$

Only the numerator in (11) is of interest since the denominator is merely a scale factor.

$$p(X_{k+1}\cdots X_{k+\Delta}|\omega_k,\theta,\underline{X}_k) = \sum_{\omega_{k+1}}\cdots\sum_{\omega_{k+\Delta}} p(X_{k+1}\cdots X_{k+\Delta}|\omega_{k+1}\cdots\omega_{k+\Delta},\theta)\, P(\omega_{k+1}\cdots\omega_{k+\Delta}|\theta,\omega_k,\underline{X}_k) \qquad (12)$$

$p(X_{k+1}\cdots X_{k+\Delta}|\omega_{k+1}\cdots\omega_{k+\Delta},\theta)$ is a known function and

$$P(\omega_{k+1}\cdots\omega_{k+\Delta}|\theta,\omega_k,\underline{X}_k) = P(\omega_{k+\Delta}|\omega_{k+\Delta-1}) \cdot\cdot\cdot P(\omega_{k+1}|\omega_k) \qquad (13)$$

$$P(\omega_k|\underline{X}_k,\theta) = p(X_k|\theta,\omega_k) \sum_{\omega_{k-1}} p(\omega_k|\omega_{k-1}) \cdot p(\omega_{k-1}|\theta,\underline{X}_{k-1})/p(X_k|\theta,\underline{X}_{k-1}) \qquad (14)$$

This is a recursive relation for $P(\omega_k|\underline{X}_k,\theta)$. The process starts from $P(\omega_1|\theta,\underline{X}_0) = P(\omega_1)$. The learning of the parameter θ can be obtained as follows:

$$P(\theta|\underline{X}_{k+\Delta}) = p(X_{k+1}\cdots X_{k+\Delta}|\theta,\underline{X}_k)\, P(\theta|\underline{X}_k)/p(X_{k+1}\cdots X_{k+\Delta}|\underline{X}_k) \qquad (15)$$

$$p(X_{k+1}\cdots X_{k+\Delta}|\theta,\underline{X}_k) = \sum_{\omega_{k+1}}\cdots\sum_{\omega_{k+\Delta}} p(X_{k+1}\cdots X_{k+\Delta}|\theta,\omega_{k+1}\cdots\omega_{k+\Delta})\, P(\omega_{k+1}\cdots\omega_{k+\Delta}|\theta,\underline{X}_k) \qquad (16)$$

$$P(\omega_{k+1}\cdots\omega_{k+\Delta}|\theta,\underline{X}_k) = P(\omega_{k+\Delta}|\omega_{k+\Delta-1}) \cdot\cdot\cdot P(\omega_{k+2}|\omega_{k+1}) \sum_{\omega_k} P(\omega_{k+1}|\omega_k)\, P(\omega_k|\theta,\underline{X}_k) \qquad (17)$$

where $P(\omega_k|\theta,\underline{X}_k)$ is taken from (14). $p(\theta|\underline{X}_k)$ can be computed recursively from:

$$P(\theta|\underline{X}_k) = \Big\{ \sum_{\omega_k}\sum_{\omega_{k-1}} p(X_k|\theta,\omega_k)\, P(\omega_k|\omega_{k-1})\, P(\omega_{k-1}|\theta,\underline{X}_{k-1}) \Big\}\; P(\theta|\underline{X}_{k-1})/p(X_k|\underline{X}_{k-1}) \qquad (18)$$

where $P(\omega_{k-1}|\theta,\underline{X}_{k-1})$ is known from (14) computed at the previous stage. Having obtained recursive solutions for $P(\omega_k|\underline{X}_k,\theta)$ and $P(\theta|\underline{X}_k)$, the computation for $P(\omega_k|\underline{X}_k,\theta)$ can be carried out by using (11), (12), and (13) and $P(\underline{\theta}|\underline{X}_{k+\Delta})$ can be computed using (15), (16), and (17). $P(\omega_k|\underline{X}_{k+\Delta})$ is then obtained from (10). The optimal adaptive decision system is again of fixed memory.

REFERENCES

1. S. C. Fralick, "Learning to Recognize Patterns without a Teacher," IEEE Trans. Information Theory, Vol. IT-13, pp. 57-64, January 1967.

2. J. Spragins, "Learning without a Teacher," IEEE Trans. Information Theory, Vol. IT-12, pp. 223-230, April 1966.

3. C. G. Hilborn and D. G. Lainiotix, "Optimal Unsupervised Learning Multicategory Dependent Hypothese Pattern Recogniztion," IEEE Trans. Information Theory, Vol. IT-14, pp. 468-470, May 1968.

4. ________, "Unsupervised Learning Minimum Risk Pattern Classification for Dependent Hypotheses and Dependent Measurements," IEEE Trans. Systems Science and Cybernetics, Vol. SSC-5, pp. 109-115, April 1969.

5. E. A. Patrick, "On a Class of Unsupervised Estimation Problems," IEEE Trans. Information Theory, Vol. IT-14, pp. 407-415, May 1968.

ACKNOWLEDGEMENT

This work has been supported by Air Force Office of Scientific Research, Office of Aerospace Research, under Contract F44620-69-C-0023.

ON THE FIELD CLASSIFICATION TECHNIQUES*

Teddy Huang
School of Electrical Engineering
Purdue University
Lafayette, Indiana 47907

ABSTRACT

Based on the concept of distance between distributions, the field classification techniques based on the Bhattacharyya distance measure, the Bayes test and the Kolmogorov-Smirnov test are formulated for both parametric and nonparametric cases. By field classification we mean that a collection of observations as a whole is classified into one of several pattern classes with known distributions. In the case of multivariate Gaussian distributions, extremely simple classification algorithms can be derived. Some field classification results of remotely sensed multispectral data are reported.

I. INTRODUCTION

In accordance with general crop planting practice, if an agricultural field has a certain majority of resolution elements classified as class ω_i, then one can feel reasonably sure that the entire resolution elements contained in the field as a whole is in class ω_i. Based on this observation, a field classification procedure is formulated as follows [2]: let $\{\omega_i, i=1,\ldots,m\}$ be a finite set of pattern classes with their distributions $\{F_i, i=1,\ldots,m\}$, respectively, and let the feature vector $\underline{x}$ be a random variable under consideration. Then given a set of observations $\{\underline{x}_i, i=1,\ldots,n\}$ all from the same field or class, the problem is to decide to which F_i they belong [5]. We shall propose several distance measures $d(F_i,F_j)$ as a suitable metric of the separation between any two distributions F_i and F_j, and we shall use this metric as a rule for decision making. Let $G(\underline{x})$ be the distribution governing a set of n observations $\{\underline{x}_1,\ldots,\underline{x}_n\}$. We shall compare the magnitudes of $\{d(F_i,G), i=1,\ldots,m\}$; the distribution F_i which minimizes the distance $\{d(F_i,G), i=1,\ldots,m\}$ is presumed to be the distribution which contains the distribution $G(\underline{x})$. Therefore, all n observations as a whole are classified into pattern class ω_i.

If the distributions concerned are unknown, the distributions $\{F_i, i=1,\ldots,m\}$ can be empirically estimated from a finite number of observations of $\{\omega_i, i=1,\ldots,m\}$, respectively. In particular, the potential function method is used to estimate the unknown distributions. In the multivariate gaussian cases, simple classification algorithms can be derived.

---

*This work was supported by NASA Grant NGL 15-005-112 and USDA Contract 12-14-100-1029(20).

II. DEFINITION OF THE DISTANCE MEASURES

Let $F_i(\underline{x})$ and $F_j(\underline{x})$ be the cumulative distribution functions (c.d.f.) of the pattern classes ω_i and ω_j, respectively, and let $p_i(\underline{x})$ and $p_j(\underline{x})$ be their probability density functions (p.d.f.) defined in Ω_x. Then the distance measure between F_i and F_j is defined as follows:

(a) Bhattacharyya distance measure [1]

$$d_B^2(F_i,F_j) = \int_{\Omega_x} \left(\sqrt{p_i(\underline{x})} - \sqrt{p_j(\underline{x})}\right)^2 d\underline{x} \ . \tag{1}$$

If we define

$$\rho_B(F_i,F_j) = \int_{\Omega_x} \sqrt{p_i(\underline{x})}\sqrt{p_j(\underline{x})} \ d\underline{x} \ , \tag{2}$$

we have

$$d_B^2(F_i,F_j) = 2 - 2\rho_B(F_i,F_j) \ . \tag{3}$$

The quantity $\rho_B(F_i,F_j)$ expresses the correlation between distributions F_i and F_j and it assumes a maximum value of 1 when F_i is identical to F_j. Therefore, we can use $\rho(F_i,F_j)$ in decision making.

(b) Kolmogorov-Smirnov distance measure [4]:

$$d_k(F_i,F_j) = \sup_{\underline{x}\in\Omega_x} |F_i(\underline{x})-F_j(\underline{x})|. \tag{4}$$

For the case that $\underline{x}$ assumes discrete values, it can be shown that

$$\sup_{\underline{x}\in\Omega_x} |F_i(\underline{x})-F_j(\underline{x})| \tag{5}$$

occurs at one of those discrete points.

III. DEFINITION OF BAYES TEST

Let the observations consist of n feature vectors: $\underline{x}_1,\ldots,\underline{x}_n$. The likelihood function $\{L_i(\underline{x}_1,\ldots,\underline{x}_n), i=1,\ldots,m\}$ is defined as follows [3]:

$$L_i = p(\underline{x}_1,\ldots,\underline{x}_n|\omega_i).$$

If the observations are independent, identically distributed random variables, the joint density is simply the product of the individual densities. The decision rule is to classify all n observations into class ω_i which maximizes the likelihood function.

IV. DISTANCE MEASURE IN NONPARAMETRIC CASES

If the distributions concerned are completely unknown, the usual assumption of being multivariate guassian distribution is not always satisfactory. If this is the case, empirically estimated distribution and density functions must be used in distance calculation defined in Sections II and III.

Distribution Function

Let $\underline{x}_1^i,\underline{x}_2^i,\ldots,\underline{x}_n^i$ be a sequence of n independent samples drawn from the distribution function $F_i(\underline{x}) = P_r[\underline{x}_j^i < \underline{x}]$ and put

$$u(\underline{x}) = \begin{cases} 1 & \underline{x} \geq 0 \\ 0 & \underline{x} < 0 \end{cases} \tag{6}$$

in which $\underline{x} \geq 0$ if and only if every component of $\underline{x}$ is greater than or equal to zero. Then the function

$$F_{ni}(\underline{x}) = \frac{1}{n} \sum_{j=1}^{n} u(\underline{x}-\underline{x}_j^i) \tag{7}$$

is called the empirical distribution function of the data. Clearly, $F_{ni}(\underline{x})$ is the proportion of the $\underline{x}_j^i$, $j=1,\ldots,n$ which are less than $\underline{x}$. Consequently, the Kolmogorov-Smirnov distance measure can be defined as follows:

$$\begin{aligned} d_K(F_i,F_j) &= \sup_{\underline{x} \in \Omega_x} |F_{ni}(\underline{x})-F_{mj}(\underline{x})| \\ &= \sup_{\underline{x} \in \Omega_x} \left| \frac{1}{n} \sum_{r=1}^{n} u(\underline{x}-\underline{x}_r^i) - \frac{1}{m} \sum_{r=1}^{m} u(\underline{x}-\underline{x}_r^j) \right| \end{aligned} \tag{8}$$

in which $(\underline{x}_1^i,\ldots,\underline{x}_n^i)$ is a sequence of n independent samples drawn from $F_i(\underline{x})$ and $(\underline{x}_1^j,\ldots,\underline{x}_m^j)$ is a sequence of m independent samples drawn from $F_j(\underline{x})$.

Probability Density Functions

The probability density function $p_i(\underline{x})$ may be empirically estimated from a sequence of n independent samples drawn from $p_i(\underline{x})$ using the potential function method. The potential function used may be of the following form [6]:

$$K(\underline{x},\underline{y}) = \exp[-(\underline{x}-\underline{y})^T(\underline{x}-\underline{y})]. \tag{9}$$

The nth estimate of the probability density function $p_i(\underline{x})$ can be expressed as:

$$p_{ni}(\underline{x}) = \frac{1}{n} \sum_{k=1}^{n} K(\underline{x},\underline{x}_k^i) \tag{10}$$

in which $\underline{x}_k^i$ is the kth sample drawn from $p_i(\underline{x})$. It has been shown that $p_{ni}(\underline{x})$ converges to $p_i(\underline{x})$ in the mean square sense. Substituting $p_{ni}(\underline{x})$ for $p_i(\underline{x})$ in eq. 2, the Bhattacharyya distance measure may be expressed as follows:

$$d_B^2(F_i,F_j) = 2 - 2\rho_B(F_i,F_j) \tag{11}$$

in which

$$\rho_B(F_i,F_j) = \int_{\Omega_x} \left[\frac{1}{n_i} \sum_{k=1}^{n_i} K(\underline{x},\underline{x}_k^i) \cdot \frac{1}{n_j} \sum_{k=1}^{n_j} K(\underline{x},\underline{x}_k^j) \right]^{1/2} d\underline{x} \tag{12}$$

V. DISTANCE MEASURE IN MULTIVARIATE GAUSSIAN CASES

Let Ω_x be a n-dimensional feature space, and let $\{F_i, i=1,\ldots,m\}$ be m multivariate guassian distributions with their probability density functions $\{p_i(\underline{x}), i=1,\ldots,m\}$ defined in Ω_x as follows:

$$p_i(\underline{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)\right], \qquad \text{for } i=1,\ldots,m \qquad (13)$$

in which $\underline{\mu}_i$ and Σ_i are the mean vector and covariance matrix of class ω_i, respectively. Substituting the expression of $p_i(\underline{x})$ defined in eq. (13) into eq. (2), we obtain the following results:

$$\rho_B(F_i,F_j) = \frac{|\Sigma_i^{-1}\ \Sigma_j^{-1}|^{1/4}}{|\frac{1}{2}(\Sigma_i^{-1}+\Sigma_j^{-1})|^{1/2}} \exp\left[-\frac{1}{4}(\underline{\mu}_i^T\Sigma_i^{-1}\underline{\mu}_i+\underline{\mu}_j^T\Sigma_j^{-1}\underline{\mu}_j) + \frac{1}{4}(\Sigma_i^{-1}\underline{\mu}_i+\Sigma_j^{-1}\underline{\mu}_j)^T(\Sigma_i^{-1}+\Sigma_j^{-1})^{-1}(\Sigma_i^{-1}\underline{\mu}_i+\Sigma_j^{-1}\underline{\mu}_j)\right]. \qquad (14)$$

Case 1. When $\Sigma_i = \Sigma_j = \Sigma$, the expression for $\rho_B(F_i,F_j)$ defined in eq. (14) may be simplified as follows:

$$\rho_B(F_i,F_j) = \exp\left[-\frac{1}{8}(\underline{\mu}_i-\underline{\mu}_j)^T\,\Sigma^{-1}(\underline{\mu}_i-\underline{\mu}_j)\right]. \qquad (15)$$

Case 2. When $\underline{\mu}_i = \underline{\mu}_j = \underline{\mu}$, the expression for $\rho_B(F_i,F_j)$ defined in eq. (14) may be simplified as follows:

$$\rho(F_i,F_j) = \frac{|\Sigma_i^{-1}\ \Sigma_j^{-1}|^{1/4}}{|\frac{1}{2}(\Sigma_i^{-1}+\Sigma_j^{-1})|^{1/2}}. \qquad (16)$$

If the sequence of r observations $(\underline{x}_1,\ldots,\underline{x}_r)$ are statistically independent, the ln of the likelihood function in multivariate gaussian cases can be expressed simply as

$$\ln L_i = -\frac{1}{2}\sum_{k=1}^{r}(\underline{x}_k-\underline{\mu}_i)^T\Sigma_i^{-1}(\underline{x}_k-\underline{\mu}_i) - \frac{1}{2}\ln|\Sigma_i| + C, \quad i=1,\ldots,m \qquad (17)$$

in which C is a constant independent of i.

VI. SYSTEM IMPLEMENTATION

Description of Data

The proposed pattern classification technique in multiclass pattern recognition has been tested on an IBM system 360/Model 44 digital computer by performing crop classification experiments. The data used in these experiments was gathered by an airborne multi-band optical-mechanical scanner. The output of this particular scanner provides 12 electrical signals, each of which is proportional to radiant energy from the scene in a different spectral band. The 12 bands cover the range from 0.4-1.0 microns in the visible and near-infrared portions of the spectrum. By simultaneously sampling the output of 12 channels, one obtains a vector which contains all the spectral information available about a given resolution element on the ground.

From the histogram study, the data can be fairly reasonably modeled by multivariate gaussian distributions with unequal covariance matrices. For the classification experiments performed, more than four different classes are considered.

General Procedure

(1) The distribution F_i for each of all pattern classes considered is estimated from an adequate number of training samples.

(2) The distribution G of an unknown pattern class is empirically estimated from a finite number of observations.

(3) Calculate the distance

$$\{d(F_i, G), \quad i=1,\ldots,m\}$$

or likelihood function

$$\{p(\underline{x}_1,\ldots,\underline{x}_n|\omega_i)\} .$$

(4) The final decision as to the classification of the set of a finite number of observations is made by choosing the class ω_i which minimizes the distance $\{d(F_i,G),\ i=1,\ldots,m\}$ or maximizes the likelihood function $\{p(\underline{x}_1,\ldots,\underline{x}_n|\omega_i),\ i=1,\ldots,m)\}$.

Experimental Results

Based on the histogram study, a number of subclasses are chosen to represent each of the pattern classes considered due to various crop types and growing conditions. For example, in area C-4, five subclasses are chosen to represent each of the following 3 classes: soybeans, corn, and pasture; four subclasses are used to represent the stubble. This is evident from the multimodal distribution of each of the four pattern classes. The statistical parameters, i.e., mean vector and covariance matrix, for each subclass are empirically estimated from a set of training samples.

In order to evaluate the performance of the field classifier, remotely-sensed multispectral data collected from agricultural fields in areas C-2, C-3 and C-4 in Tippecanoe County, Indiana is tested based on field classification algorithms defined in this paper. The classification results obtained by employing both Field Classification Algorithms and the Maximum Likelihood Decision Rule (MLDR) are shown in Table 1-Table 5. The performance of the field classifier is defined as the percentage of correctly classified fields in the area considered while that of the MLDR classifier is defined as the percentage of correctly classified resolution elements in the same area. For example, there are 81 agricultural fields which consist of 22,353 resolution elements in area C-4. According to the experiment results shown in Table 1, 75 fields which consist of 21,347 resolution elements are correctly classified by the field classifier while 17,987 resolution elements are correctly classified by the MLDR classifier. The performance of the field classifier and the MLDR classifier are 92.6 and 79.7, respectively. Hence, the performance of the field classifier is considerably better than that of the MLDR classifier. It is also evident from these experiments that the computer time required by the field classifier for making a decision about the classification of a set of data is considerably less than that of the MLDR classifier.

VII. CONCLUSION

As far as the multispectral agricultural data is concerned, the performance of the proposed field classification techniques are considered to be quite satisfactory. Nonparametric methods are required when the distributions for all the pattern classes concerned are unknown. In the multivariate gaussian cases, the field classification algorithms are considerably simpler and more efficient than that of the maximum likelihood classifier.

ACKNOWLEDGMENT

The author wishes to express his appreciation to Professor K. S. Fu, Professor D. A. Landgrebe and Dr. P. H. Swain for their many helpful suggestions and discussions about the work reported here. The crop classification experiments were done at the Laboratory for Applications of Remote Sensing, Purdue University.

REFERENCES

1. Matusita, K., "Classification Based on Distance in Multivariate Gaussian Cases," Fifth Berkeley Symposium on Math. Stat. and Prob., Vol. 1, 1965-66, pp. 299-304.

2. Huang, T., "Per-field Classifier for Agricultural Applications," LARS Information Note 042369, Purdue University, 1969.

3. Van Trees, H., Detection, Estimation, and Modulation Theory I, Wiley, 1968.

4. Darling, D. A., "The Kolmogorov-Smirnov, Cramer-Von Mises Test," Annals of Math. Stat., Vol. 28, 1957, pp. 823-838.

5. Ellison, B. E., "A Classification Problem in which Information About Alternative Distributions is Based on Samples," Annals of Math. Stat., Vol. 33, 1962, pp. 213-223.

6. Lemke, R. R., "On the Application of the Potential Function Method to Pattern Recognition and System Identification," TR-EE 68-8, Purdue University, 1968.

TABLE 1

CLASSIFICATION RESULTS OF AIRCRAFT DATA TAKEN FROM AREA C-4 IN JULY, 1966, BASED ON B-DISTANCE

Class	No. of Fields	No. of Fields Classified Into: Soybean	Corn	Pasture	Stubble
Soybean	19	17	2	0	0
Corn	31	1	30	0	0
Pasture	18	0	0	18	0
Stubble	13	0	0	3	10

Total No. of Fields Classified = 81
No. of Fields Correctly Classified Based on B-Distance Measure = 75
Performance of Field Classifier = 92.6
Performance of MLDR Classifier = 79.7

TABLE 2

CLASSIFICATION RESULTS OF AIRCRAFT DATA TAKEN FROM AREA C-3 IN JULY, 1966, BASED ON B-DISTANCE

Class	No. of Fields	No. of Fields Classified Into: Soybean	Corn	Mixture	Water
Soybean	13	11	1	1	0
Corn	10	0	9	1	0
Mixture*	18	1	0	17	0
Water	3	0	0	0	3

Total No. of Fields Classified = 44
No. of Fields Correctly Classified Based on B-Distance Measure = 40
Performance of Field Classifier = 90.9
Performance of MLDR Classifier = 76.6

*Mixture - Mixture of pasture and stubble

TABLE 3

CLASSIFICATION RESULTS OF AIRCRAFT DATA TAKEN FROM AREA C-2 IN SEPTEMBER, 1966, BASED ON B-DISTANCE

Class	No. of Fields	No. of Fields Classified Into: Soybean	Corn	Pasture	Stubble	Water
Soybean	12	8	4	0	0	0
Corn	14	1	13	0	0	0
Pasture	13	0	0	13	0	0
Stubble	11	0	1	1	9	0
Water	3	0	0	0	0	3

TABLE 3 (cont.)

Total No. of Fields Classified = 53
No. of Fields Correctly Classified Based on B-Distance Measure = 46
Performance of Field Classifier = 86.8
Performance of MLDR Classifier = 67.0

TABLE 4

CLASSIFICATION RESULTS OF AIRCRAFT DATA TAKEN FROM AREA C-4 IN JULY, 1966, BASED ON BAYES TEST

Class	No. of Fields	No. of Fields Classified Into: Soybean	Corn	Pasture	Stubble
Soybean	19	17	2	0	0
Corn	31	3	28	0	0
Pasture	18	2	0	14	2
Stubble	13	0	0	3	10

Total No. of Fields Classified = 81
No. of Fields Correctly Classified Based on Bayes Test = 69
Performance of Field Classifier = 85.5
Performance of MLDR Classifier = 79.7

TABLE 5

CLASSIFICATION RESULTS OF AIRCRAFT DATA TAKEN FROM AREA C-2 IN SEPTEMBER, 1966, BASED ON BAYES TEST

Class	No. of Fields	No. of Fields Classified Into: Soybean	Corn	Pasture	Stubble	Water
Soybean	12	8	4	0	0	0
Corn	14	2	12	0	0	0
Pasture	13	0	0	13	0	0
Stubble	11	0	1	5	5	0
Water	3	0	0	0	0	3

Total No. of Fields Classified = 53
No. of Fields Correctly Classified Based on Bayes Test = 41
Performance of Field Classifier = 77.5
Performance of MLDR Classifier = 67.0

A PATTERN RECOGNIZING STUDY OF PALM READING *

MORIYA ODA**
Automatic Control Laboratory, Faculty of Engineering, Nagoya University, Nagoya, Japan
BAXTER F. WOMACK
Department of Electrical Engineering, Electronics Research Center, The University of Texas at Austin, Austin, Texas 78712
KENJI TSUBOUCHI
Lamp & Electronic Component Section,
Mitsui & Co., Ltd., Tokyo, Japan

Abstract — This paper studies "palm reading" from the viewpoints of pattern recognition, heuristics and learning. Two kinds of mapping in palm reading are extracted, these being a tree-decision type and a polynomial type. In contrast to the fairly simple form of mapping, the palm input is relatively complicated but well chosen. This paper deals mainly with the generalization and speciality of palm reading, variations of palm lines, examples of palmistry judgment, and the roles of heuristics and learning in palm reading.

1. INTRODUCTION

The purpose of this paper is to clarify the decision mechanism of palm reading (palmistry, or chiromancy) from the viewpoints of pattern recognition, learning, and heuristics. Palm reading can be considered as a special problem in what is widely-defined as pattern recognition. The study of this problem seems very important but is difficult and therefore similar to the problems of reading hand-written characters and numerals, and of identifying human faces.

Generally, we can list the following three kinds of approaches to the study of the hand:

(1) Medico-legal study of fingerprints and palms*1 as a means of identification [1],

(2) Dermatoglyphical study of fingerprints, palms and soles, which includes anthropology, medicine, genetics and anatomy [2,3], and

(3) Occult study of hands (fingers and palms) as in palmistry or chiromancy [4 - 9], which requires knowledge in modern psychology, anatomy and phylogeny [9].

This paper does not concern the first two approaches given above which have been studied in a scientific way, but concentrates on the last one which seems to have been studied spuriously only in an occult way. This paper will, however, consider pattern recognizing (instead of occult) approaches to the human's decision-making mechanism of palm reading [4 - 9]. Such a study is made first in this paper according to the authors' knowledge. (This is a modified and developed version of Ref. [10].)

---

* This research was supported in part by a Post-Doctoral Fellowship and the Joint Services Electronics Program.

** Now visiting research professor at The University of Texas at Austin.

*1 As a means of identification, palm lines are better than fingerprints, but it is less practical (more difficult) to make good impressions.

Palm lines might seem complicated at a glance. They are considered sometimes to be occult or to need a special master skill for their reading. It is found, however, as a main result of this paper that most of the decision mechanism of basic palm reading can be formulated in a simple form of either (a) *tree decision* or (b) *polynomial function* of well-chosen input terms (palm lines). We will see its details in the following sections.

2. GENERAL CONSIDERATION OF PALM READING

In this section, a general consideration of palm reading is developed from a viewpoint of pattern recognition.

2.1 Palm Reading System

Fig. 1 shows a palm reading system, which includes a palm reader (palmist or palm reading machine - Block B), his subject S_i (Block A), a reference model (the same subject, the same palmist, or some other person who knows the subject well - Block D), and an adaptive loop for modifying Block B (Block C).

The four blocks in Fig. 1 might be covered in several ways, for instance (1) by a single person who is a palmist as well as a subject; or (2) by a palmist playing the part of Block B, or Blocks B and C, or Blocks B, C and D and his subject correspondingly playing the part of Blocks A, D and C, or Blocks A and D, or Block A; or (3) by a palmist, his subject and a third person who knows the subject well.

2.2 Mapping in Palm Reading

Let's consider the block B in Fig. 1. The input (vector) ${}_i\underline{X} \in X$ (set of inputs) corresponds to the palm lines (simply *palm*) of the subject S_i in a narrow sense or the hand (including fingers and palm) in a wide sense. A palm reader (generally palmist) translates the input ${}_i\underline{X}$ into an output (vector) ${}_i\underline{Y} \in Y$ (set of outputs), which consists of subvectors (subsets) ${}_i\underline{Y}_j$'s $(j=1, 2, \cdots)$; ${}_i\underline{Y}_j = ({}_iY_{jk}) = ({}_iY_{j1}, {}_iY_{j2}, \cdots)$, where the subvector ${}_i\underline{Y}_j$ corresponds to a *palmistry item* such as intelligence, emotional caliber, physical energy, temperament compatibility, married life, occupational life, life success, etc., and an element ${}_iY_{jk}$ which corresponds to a *palmistry term* (character) such as high or low intelligence, good or bad mental concentration, sympathy or egoism, practicality or imagination; warm or cold personality, dependence or independence, good or bad emotional concentration, restless emotionality, strong or weak constant search for new experiences, strong or weak inner tension, criminal (destructive) or creative tendencies; high or low energy potential, strong or weak ambitious motivation, stable or unstable physical energy; and so on, and ${}_iY_{jk} = 1$ when such a character is judged to exist, and ${}_iY_{jk} = 0$ when judged not to exist.

Therefore, we can formulate the problem of Block B in Fig. 1 as a mapping from X-space (space of palms) to Y-space (space of palmistry judgment). See Fig. 2.

2.3 Two Types of Mapping

Through checking several books on palm reading [4 - 9], the authors have found there are two main kinds of mapping M in palm reading as follows:

(a) Tree decision type mapping from $\underline{X}$ to $\underline{Y}$, defined by

$$\underline{Y} = f_t(\underline{X}) \tag{2.1}$$

(b) Polynomial type mapping from $\underline{X}$ to $\underline{Y}$, defined by

$$\underline{Y} = f_p(\underline{X}), \tag{2.2}$$ [*2]

where $\underline{X}$ and $\underline{Y}$ are used here and hereafter for the convenience of simplicity instead of ${}_i\underline{X}$ and ${}_i\underline{Y}$, respectively.

In the actual problem of palm reading, the components X_{jk}'s of $\underline{X}(j=1,2,\cdots; k=1,2,\ldots)$ in $f_t(\underline{X})$ and $f_p(\underline{X})$ are selected properly but differently by each palmist. We will consider several examples later in Section 3.5.

2.4 Specialities in Mapping in Palm Reading

There are several specialities in palm reading which are concerned directly or indirectly with the mapping M.

2.4.1 Growth of a Person and Change of his Palm. It is said that the palm (lines) $\underline{X}$ of a person changes very slowly but steadily with age, and also the subject grows with age—therefore, the palmistry judgment $\underline{Y}$ for the subject should be changed correspondingly. But, there is a difference between the changing rates of the subject and his palm. This is one of the problems which makes palmistry judgment rather difficult.

2.4.2 Incomplete Partition of Palm Judgment Space. In Sec. 2.2, the palm judgments $\underline{Y}_j$'s and Y_{jk}'s (left subscript i is omitted) are defined as palmistry items and terms, respectively. There is a possibility of overlaps (intersections) between $\underline{Y}_j$ and $\underline{Y}_{j'}$ $(j \neq j')$, Y_{jk} and $Y_{jk'}$ $(k \neq k')$, or Y_{jk} and $Y_{j'k'}$ $(j \neq j')$, i.e.,

$$\left.\begin{array}{l} \underline{Y}_j \cap \underline{Y}_{j'} \neq \emptyset, \\ Y_{jk} \cap Y_{jk'} \neq \emptyset, \\ Y_{jk} \cap Y_{j'k'} \neq \emptyset. \end{array}\right\} \tag{2.3}$$

Any overlap is due to the incompleteness of the expression of human character. The overlap is a fatal (inevitable) defect from one viewpoint of pattern classification, but may be useful as a kind of redundancy in the classification of human character.

Besides, we can cite the following possibility:

$$\bigcup_k Y_{jk} \neq \underline{Y}_j, \quad \bigcup_k Y_{jk} \subset \underline{Y}_j, \tag{2.4}$$

$$\left.\begin{array}{l} \bigcup_j \underline{Y}_j \neq \text{complete description of a subject} \\ \bigcup_j \underline{Y}_j \subset \text{complete description of a subject} \end{array}\right\} \tag{2.5}$$

Eq. (2.4) means that a palmistry item $\underline{Y}_j$ may not be completely covered by existing palmistry terms Y_{jk}'s. Similarly, Eq. (2.5) means that a complete description of a subject may not be attained by existing palmistry items $\underline{Y}_j$'s.

2.4.3 Differences in Palmists. There are many differences in palmists' judgments. The main differences, which have not been unified yet, appear in selections of input palm $\underline{X}$, output judgment $\underline{Y}$, and mapping M — the form of $f_t(\underline{X})$ or $f_p(\underline{X})$ and the weights in them.

[*2] Generally, the polynomial is a first-order linear polynomial.

2.4.4 Number of Components and Their Combinations of Input Palm.

Generally, the number of components of input palm $\underline{X}$ which is used in most palmistry judgments Y_{jk}'s is less than 7~8, which is enough to make more than $128(=2^7) \sim 256(=2^8)$ combinations of the components since they have two or more variations. (See Sec. 3.3.)

3. BASIC KNOWLEDGE AND EXAMPLES OF PALM READING

The general consideration in Sec. 2 will be described in more detail in this section.

3.1 Hand Analysis and Palm Reading

In palmistry, a hand analysis is seen more practically than palm reading. For instance, in hand analysis [4 - 9], not only the palm but also all portions of the hand (back, front, fingers, nails, skin, color, etc.) are carefully examined under a special priority within these items, which can be called heuristic priority [11 - 12], but is different for each palmist. We will, however, see how to read only an active palm, not a passive palm[*3] [9] for convenience and simplicity without any loss of generality.

3.2 Names of Palm Lines

As shown in Fig. 3, there are generally thirteen kinds of palm lines[*4], namely, Head Line ($\underline{X}_1$), Life Line ($\underline{X}_2$), Heart Line ($\underline{X}_3$), Fate Line ($\underline{X}_4$), Marriage Lines ($\underline{X}_5$), First Influence Lines ($\underline{X}_6$), Second Influence Lines ($\underline{X}_7$), Apollo Line ($\underline{X}_8$), Health Line ($\underline{X}_9$), Mars Line ($\underline{X}_{10}$), Girdle of Venus ($\underline{X}_{11}$), Ring of Solomon ($\underline{X}_{12}$), and Ring of Saturn ($\underline{X}_{13}$). We cannot find all the above lines in every palm, but always some of them.

3.3 Variational Factors of Palm Lines

There are a lot of variational factors which produce many variations X_{jk}'s $(j=1,2,\cdots, 13;\ k=1,2,\cdots)$ of the above palm lines $\underline{X}_j$'s. The factors (see Fig. 4) are spot, circle, triangle, square, grille, cross, star, island, chain, tassel, fork, sister, ascending or descending branches, wave, break, capillary, starting and ending points, direction, size and position to other lines(s), length, straightness or curvature, deepness, thickness, clearness, and so on. Some of these variations are shown in Fig. 4.

3.4 Heuristic Priority in Palm Reading

In most cases of palm reading a judgment Y_{jk} is made by considering the variations X_{jk}'s of only one palm line $\underline{X}_j$. This means that the line $\underline{X}_j$ under consideration has an extremely high priority within those lines, while all the rest have a negligible priority. In another case where more than two palm lines are regarded, there is a specific priority among them ($\underline{X}_j$'s or X_{jk}'s). Such examples are illustrated in the next section.

3.5 Examples of Palm Reading

In this section, we will see two interesting but somewhat complex examples of palm reading [9]: one is a tree decision type, and the other a first-order linear polynomial type.

[*3] An active palm corresponds to a palm with more complicated palm lines, while a passive palm is less complicated [9].

[*4] Palm lines are divided into two kinds, Main Lines and Subsidiary Lines. But, the number and kinds of main lines vary with each palmist.

3.5.1 Variations of Head, Life and Heart Lines. Before seeing the examples, we need to classify the variations of three of the main lines, i.e., Head, Life and Heart [9].

The Head Line has three variations (H, O and L type) due to the height of its ending point as shown in Fig. 5, where H type corresponds to the Heart Line ending in the 4th zone, O type in the 3rd zone, and L type in the 2nd and 1st zones.

The Life Line has six variations (A, B, ···, F type) due to the relative position among the starting-points of the Life, Head and Heart Lines as shown in Fig. 6, where A type corresponds to the Life Line connecting with the Head Line at their starting points, B type non-connecting, C type being extremely apart, D type overlapping with the Head Line, E type crossing with the Head Line, and F type overlapping with the Head and Heart Lines.

The Heart Line has three variations (Types I, II and III) due to the relative position between its ending point and two fingers, Jupiter (forefinger) and Saturn (middle finger) as shown in Fig. 7, where type I corresponds to the Heart Line ending between the forefinger and middle finger, type II under the middle finger, and type III under the forefinger.

3.5.2 Judgment of Intelligent Ability*5. Two kinds of palm inputs, Head Line $\underline{X}_1$ and Life Line $\underline{X}_2$, can decide three kinds of palmistry outputs of intelligent abilities as shown in Table 1 [9]. The table shows an example of a tree decision type $f_t(\underline{X})$ in palmistry judgment, where the first branches (or heuristic internodes [11,12]) correspond to three variations of Head Line, and each of the second branches the six variations of Life Line.

3.5.3 Judgment of Temperament Compatibility in Both Sexes*6. First, four kinds of male and female subjects' palm inputs, Life Line $\underline{X}_2$, Heart Line $\underline{X}_3$, Marriage Lines $\underline{X}_5$, and Second Influence Lines $\underline{X}_7$, calculate the scores of the first-order linear polynomial function, $f_p(\underline{X}) = f_p^M(\underline{X})$ for the male subject or $f_p^M(\underline{X})$ for female:

$$f_p(\underline{X}) = \sum_{k=1}^{6} a_{2k} X_{2k} + \sum_{k=1}^{3} (a_{3k} X_{3k} + a_{5k} X_{5k} + a_{7k} X_{7k}) \quad (3.1)$$

where

$$X_{jk} = \begin{cases} 1 & \text{if the corresponding variation exists,} \\ 0 & \text{if the corresponding variation does not exist,} \end{cases} \quad (3.2)$$

and a_{jk}'s are the corresponding weight values (cf. Table 2).

Next, the difference $\Delta f_p(\underline{X}) = f_p^F(\underline{X}) - f_p^M(\underline{X})$ can judge the subject's temperament compatibility as follows:

$$\left.\begin{array}{l} |\Delta f_p(\underline{X})| < \theta_1 \cdots \text{ good compatibility,} \\ |\Delta f_p(\underline{X})| > \theta_2 \cdots \text{ bad compatibility,} \end{array}\right\} \quad (3.3)$$

where for instance the thresholds are $\theta_1 = 20$ and $\theta_2 = 30$ in Ref. [9].

4. DISCUSSION

In this section, we will discuss more specifically the two phases of heuristics and learning in palm reading, since we've considered thus far mainly the phases where the following processes were established.

---

*5,*6 The result of palmistry judgment is not so important, but the process (or mechanism) is most important in this paper.

See Fig. 1 again.

From a viewpoint of heuristics, the input terms X_{jk}'s of mapping $f(\underline{X})$ (Table 1 or Eq. (3.1)) mean heuristic elements (or clues), and the weights a_{jk}'s represent heuristic priority (or system) [11,12].

From a viewpoint of learning, the organizing process of the decision tree, the converging process of the weights a_{jk}'s, and the searching process of properly-chosen terms X_{jk}'s in the polynomial function are all corresponding to learning processes.

5. CONCLUSIONS

This paper clarifies the decision mechanism of basic palm reading, which had seemed rather occult and one of the most human-like (intelligent) decision problems. The mapping in the problem is generally formulated in fairly simple forms such as a tree type and a first-order linear polynomial type, while the palm input is found to consist of well-chosen independent palm lines and their ingeniously-devised variations. Also in the paper are first studied the generalization and speciality of palm reading from a viewpoint of pattern recognition, examples of palm reading, and the roles of heuristics and learning in palm reading.

This paper is another step by the authors in studying interesting mechanisms with so-called human-like (not machine-like) decision. The paper will hopefully be an active forum in expanding pattern-recognition (or artificial intelligence) research into complicated problems where potential solutions can be formulated in the manner developed in this paper.

Acknowledgment. The authors would like to thank Professor K. Nakamura, Nagoya University, Japan for his kind and helpful discussions.

REFERENCES

[1] A. T. Field: Fingerprint Handbook, C. C. Thomas Publisher (1959).

[2] H. Cummins, C. Midlo: Finger Prints, Palms and Soles — An Introduction to Dermatoglyphics, Dover Publications (1961).

[3] L. S. Penrose: Dermatoglyphics, Scientific American 221 - 6, 72/74 and 79/84 and 152 (December 1969).

[4] W. G. Benham: The Laws of Scientific Hand Reading, Duell, Sloan and Pearce (Revised ed., 1966).

[5] F. Gettings: The Book of the Hand, Paul Hamlyn, Ltd., London (1965).

[6] F. Gettings: Palmistry Made Easy, Wilshire Book Co. (1966).

[7] Cheiro: Cheiro's Language of the Hand, ARC Books, Inc. (1968).

[8] E. Niles: Palmistry, HC Publishers, Inc. (1969).

[9]* H. Asano: Palmistry, Kappa Books (1962).

[10]* K. Tsubouchi: Pattern Recognizing Study in Palmistry, Department of Electrical Engineering, Nagoya University, B.S. Thesis (March 1970)-unpublished.

---

* Written in Japanese.

[11] M. Oda and B. F. Womack: A Unified Discussion on Heuristics in Artificial Intelligence, Proceedings of Third Asilomar Conference, 387/400 (December 1969).

[12] M. Oda and B. F. Womack: A Detailed Investigation of Various Heuristic Approaches in Artificial Intelligence, 1970 SWIEEECO Record, 125/129 (April 1970).

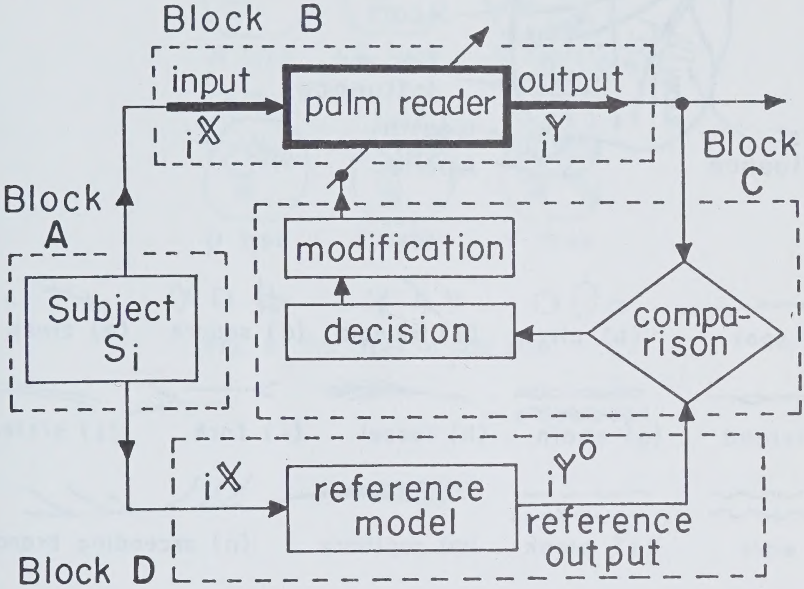

Block B
input
palm reader
output
$_i\mathbb{X}$
$_i\mathbb{Y}$
Block C
Block A
modification
Subject S_i
decision
compa-rison
$_i\mathbb{X}$
reference model
$_i\mathbb{Y}^0$
reference output
Block D

Fig. 1 Palm reading system.

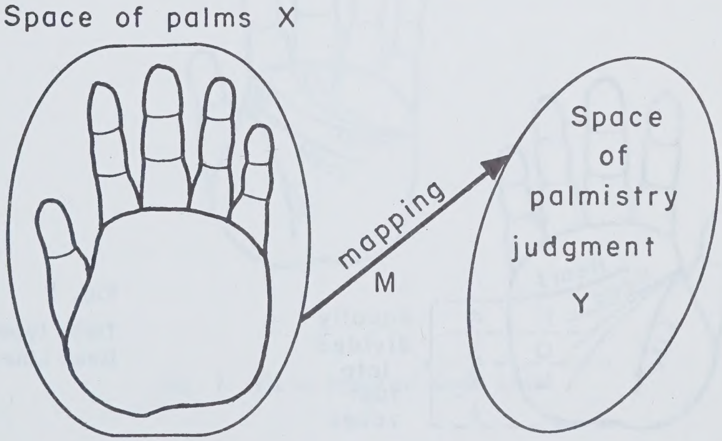

Space of palms X
mapping
M
Space of palmistry judgment Y

Fig. 2 Mapping M from X to Y.

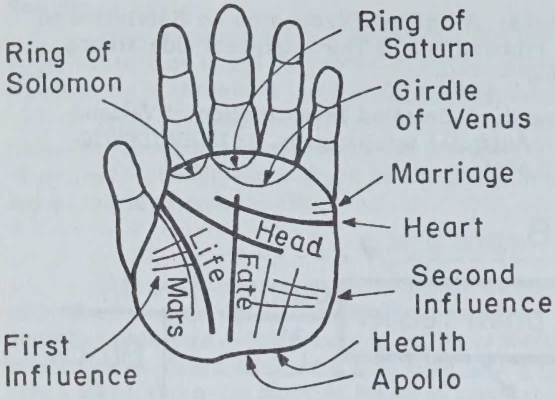

Ring of Solomon
Ring of Saturn
Girdle of Venus
Marriage
Heart
Head
Life
Fate
Mars
Second Influence
First Influence
Health
Apollo

Fig. 3

Names of palm lines.

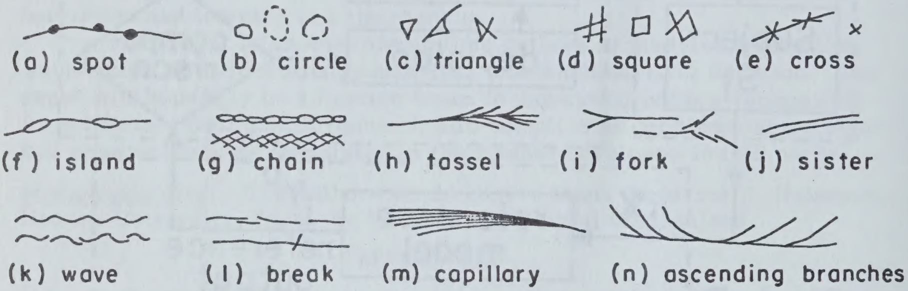

(a) spot
(b) circle
(c) triangle
(d) square
(e) cross
(f) island
(g) chain
(h) tassel
(i) fork
(j) sister
(k) wave
(l) break
(m) capillary
(n) ascending branches

Fig. 4 Variational factors of palm lines.

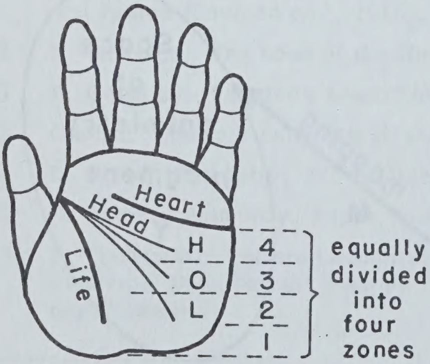
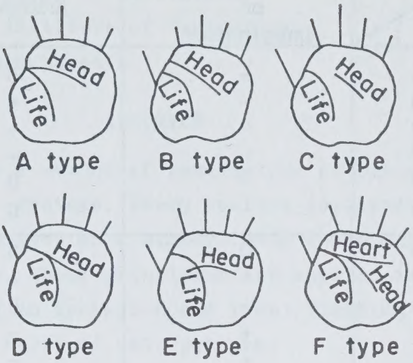

Heart
Head
Life
H
O
L
4
3
2
1
equally divided into four zones

Fig. 5

Three types of Head Line.


Head
Life
Heart

A type B type C type

D type E type F type

Fig. 6 Six types of Life Line.

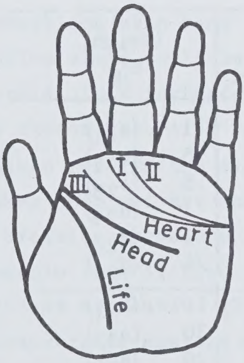

I
II
III
Heart
Head
Life

Fig. 7 Three types of Heart Line.

Table 1 Type of Palm vs. Intelligent Ability

Palm input Head Line	ability Life Line	Creation or Imagination	Arithmetic	Logic
H (X_{11})	A (X_{21})	-	0	+
	B (X_{22})	-	+	+
	C (X_{23})	-	+	+
	D (X_{24})	-	+	0
	E (X_{25})	-	0	0
	F (X_{26})	-	0	0
O (X_{12})	A (X_{21})	0	0	0
	B (X_{22})	-	+	+
	C (X_{23})	-	0	0
	D (X_{24})	+	-	0
	E (X_{25})	+	-	-
	F (X_{26})	0	-	-
L (X_{13})	A (X_{21})	+	-	-
	B (X_{22})	+	0	0
	C (X_{23})	0	0	-
	D (X_{24})	+	-	-
	E (X_{25})	+	-	-
	F (X_{26})	+	-	-

(Notice) + = high, 0 = average, - = low

Table 2 Weight Values a_{jk}'s in Eq. (3.1)

Palm input $\underline{X}_j$	X_{jk}	Weight a_{jk}
Life Line $\underline{X}_2$	A (X_{21})	1 (a_{21})
	B (X_{22})	5 (a_{22})
	C (X_{23})	5 (a_{23})
	D (X_{24})	1 (a_{24})
	E (X_{25})	1 (a_{25})
	F (X_{26})	25 (a_{26})
Heart Line $\underline{X}_3$	I (X_{31})	10 (a_{31})
	II (X_{32})	30 (a_{32})
	III (X_{33})	20 (a_{33})
Marriage Lines $\underline{X}_5$	less than 3 (X_{51})	0 (a_{51})
	more than 4 (X_{52})	5 (a_{52})
	long (X_{53})	10 (a_{53})
Second Influence $\underline{X}_7$	none (X_{71})	0 (a_{71})
	less than 5 (X_{72})	5 (a_{72})
	more than 6 (X_{73})	10 (a_{73})

PATTERN PROCESSING USING MULTILEVEL SYSTEMS

GOESTA H. GRANLUND [+]
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts.

ABSTRACT

The paper describes a method of realisation of certain multilevel pattern recognition systems. Every feature isolated on one level is checked for plausibility on a higher level where additional information is available. These principles are applied in a computer program which makes an isolation and identifikation of certain cell contours in a stylized cell pattern.

I. INTRODUCTION

A simple, but frequently used, system model for pattern recognition consists of two blocks: A feature detection block followed by a categorizer. An underlying assumption is that the feature isolating block can generate variables suitable as input to a "more or less linear" decision system. Features satisfying this requirement are difficult to find, and this depends to some extent on the fact that these two systems are discontinuous transformations. A nonspecialized pattern recognition system can be viewed in more detail as a series of discontinuous transformations with some interactions between blocks. These transformation steps could be said to represent different abstraction levels. In a textprocessing system, for example, suitable levels could be chosen according to fig. 1.

But in all practical pattern recognition systems, the cost of feature measurement has to be taken into account and this often leads to some form of sequential decision procedure. [1] The choice of suitable feature measurements is a difficult problem. Optimization with the mathematical methods available [2] is often difficult in practical cases where necessary parameters are unknown.

This paper describes an approach to isolate important features for a

---

[+] This paper describes the results of a work carried out at Chalmers University of Technology, Department of Network Theory, Gothenburg, Sweden.

specific pattern class. Its two characteristic features are that first, structural elements are isolated in two different ways and later compared by means of a redundancy or plausibility check. And second, a trial-and-error procedure is used by which the program partitions the pattern so as to begin with the easiest parts first.

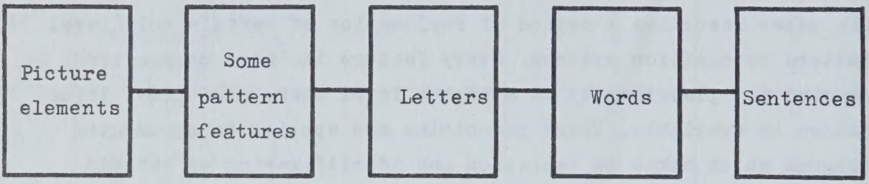


Figure 1. Example of different levels of abstraction.

II. MODEL FOR FEATURE SELECTION

The system described operates with three abstraction levels. It first tries to isolate a certain feature on the lowest abstraction level. It makes a check of the plausibility of the feature considered in isolation with respect to other features. If this check is negative the information related to the isolated feature is dropped and the system proceeds to another part of the feature space at the lowest level.The procedure containing a check for plausibility together with trial and error forms a higher level of abstraction. The assumption underlying this approach is that it is necessary to synthesize a certain feature on a specific level, but since all the necessary information is not available on that level, the feature isolated might appear to be insignificant later. A check can be made against other features, and this gives a relation to properties containing more information. As the synthesis proceeds these checks can be made with more significance. Fig. 2 is an attempt to make a generalisation of the procedure described. To make the scheme clearer, the features used in this application are specified. The first (lowest) level is not explicit in this scheme as the first synthesis step deals with features to be checked on the second level. However, a practical system needs some start-up procedure where the information of the first level has to be taken into account. This part will be discussed in

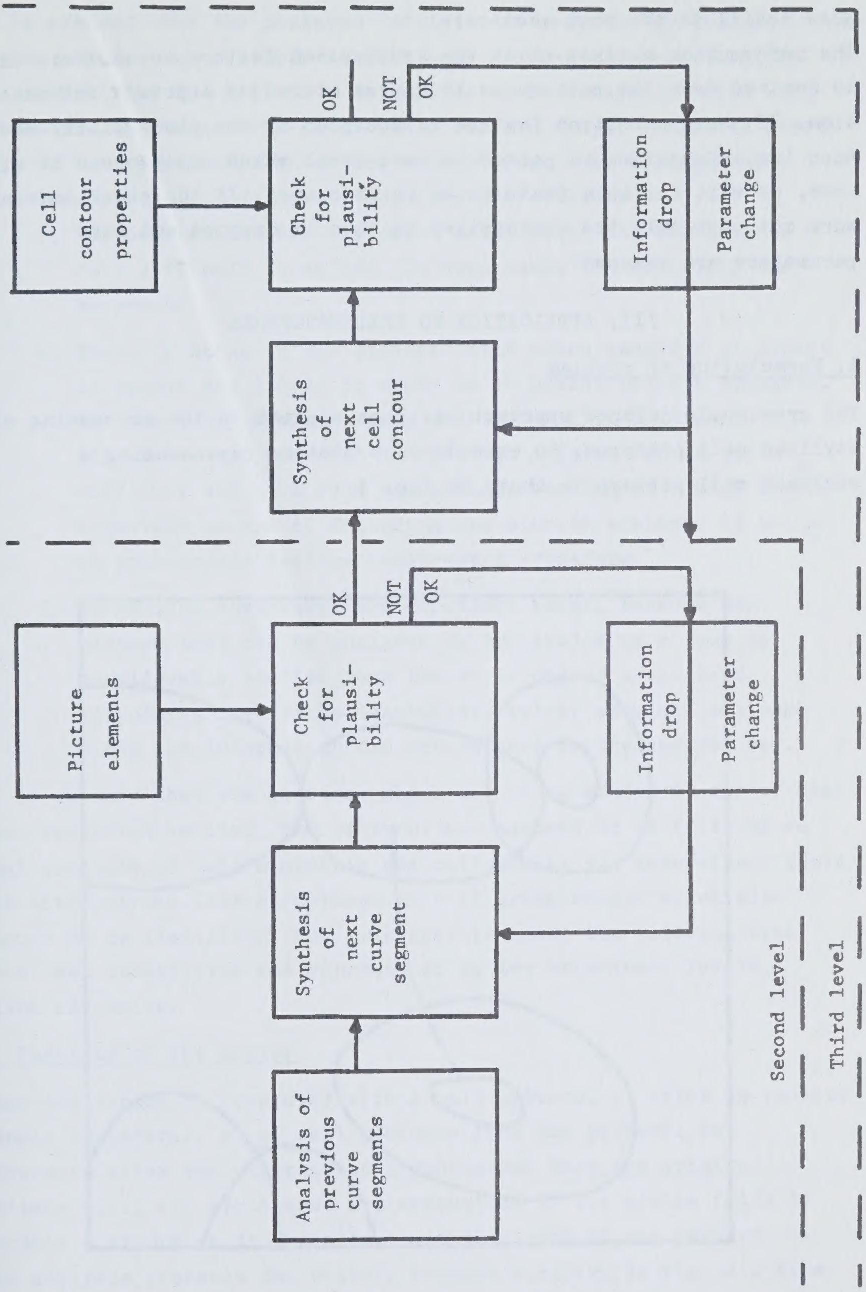

Analysis of previous curve segments
Synthesis of next curve segment
Picture elements
Check for plausi-bility
OK
NOT OK
Information drop
Parameter change
Second level
Synthesis of next cell contour
Cell contour properties
Check for plausi-bility
OK
NOT OK
Information drop
Parameter change
Third level

Figure 2. Generalised function diagram.

more detail in the next section.
The information against which the synthesized feature is checked, can be derived from the pattern or it can be stored as a priori information. If the synthesized feature is accepted by the plausibility check, then the information is passed to next level which may process it at once, or wait for more features to be extracted. If the check does not work out correctly the synthesized feature is dropped and some parameters are changed.

III. APPLICATION TO CELLSTRUCTURES

A. Formulation of problem

The previously defined approach has been applied to the processing of stylized cell patterns. An example of a picture representing a stylized cell pattern is shown in fig. 3.

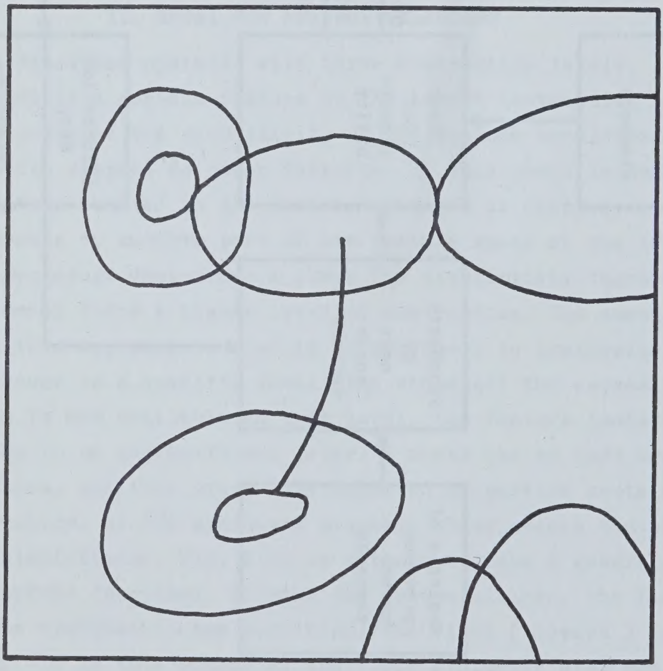


Figure 3. Example of a stylized cell pattern.

It is assumed that the patterns fed into the computer system are differentiated and filtered so that only the contours are left. We can assume that the goal is a pattern recognition system which will detect the existence of a cell with a certain shape. Yet it is difficult to find parameters which are characteristic of a single cell. Thus from fig. 3 the following problems are evident:

1. The overlap between different cells makes the boundary of a cell difficult to define. In some cases the ambiguity can be great.
2. There is noise in the picture. This noise consists of breaks in curves and lines, or spots on or beside pattern elements. The question of what breaks are insignificant and can be filtered away, and what is relevant information is quite difficult and as a rule requires processing at a high level. A certain amount of filtering can also be achieved by using an appropriate feature measurement procedure.
3. Incomplete structures are sometimes found, because the picture that can be analysed by the system as a rule is considerably smaller than the whole object area. Such structures have to be identified, but as they are not complete the information can not be used in the desired way.

It is assumed that the pictures which are to be analysed, are of the form described earlier. The pictures are assumed to be filtered so that contours of cell membranes and cell nuclei are emphasized. There are often thread-like structures in cell preparations which also should be be identified. For practical reasons, the patterns used have been constructed and represented by two brightness levels, black and white.

B. Function of the system

When the system is presented with a cell pattern, it tries to isolate single structures, e. g. cell contours from the pattern. One structure after the other is then subtracted from the original pattern until all structures are exhausted. If the system fails to isolate a structure it proceeds to another part of the pattern. As the analysis proceeds the pattern becomes simpler. In fig. 4 a flow diagram of the system is shown.

The pattern read into the computer consists of a 35 x 35 binary matrix. The system first scans the pattern in a search for the first

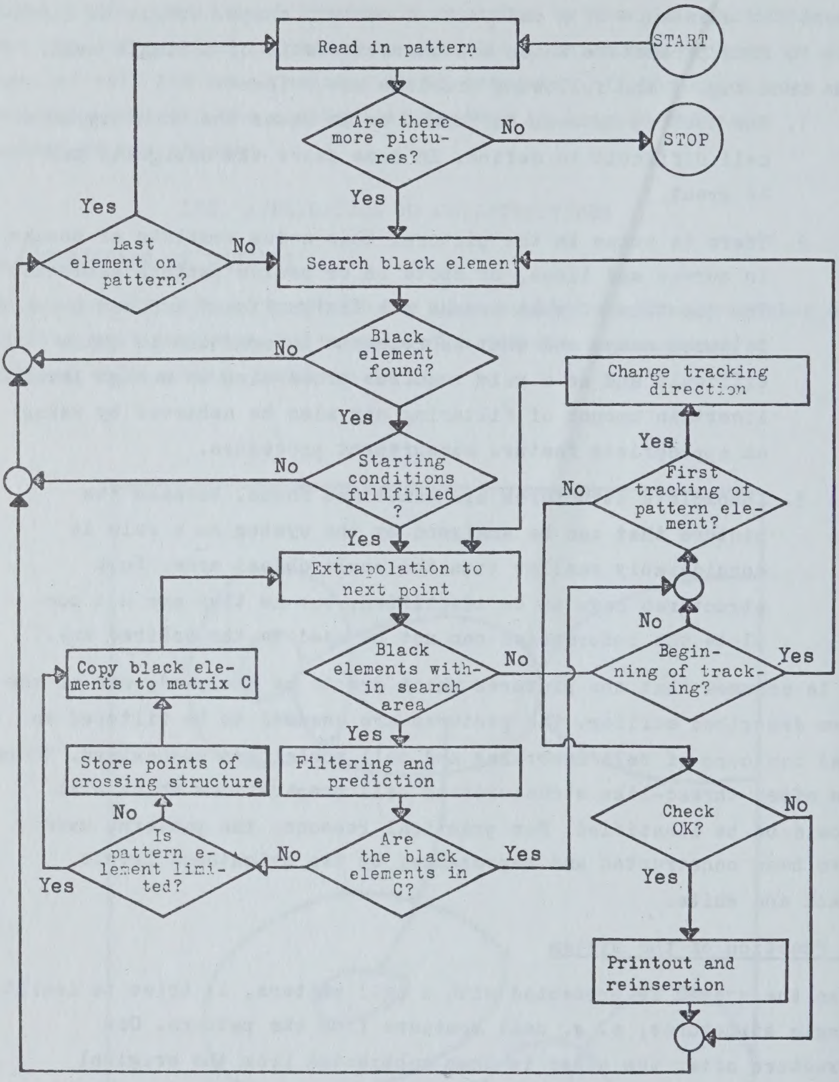

START
Read in pattern
Are there more pictures?
No
STOP
Yes
Last element on pattern?
Yes
No
Search black element
Black element found?
No
Yes
Starting conditions fullfilled ?
No
Yes
Extrapolation to next point
Black elements within search area
No
Yes
Filtering and prediction
Are the black elements in C?
No
Yes
Is pattern element limited?
No
Yes
Store points of crossing structure
Copy black elements to matrix C
Change tracking direction
Yes
First tracking of pattern element?
No
No
Beginning of tracking?
Yes
Check OK?
No
Yes
Printout and reinsertion

Figure 4. Flow diagram of system program.

black element. Before the system starts to track the structure, it makes some checks around the black element found in order to ascertain that the first part of the structure goes in the same direction as the scanning.

The system then makes an extrapolation to the next point on the structure. The data fed into the extrapolation procedure consists of filtered data derived from previous points on the structure, or is the scanning direction if the tracking of the structure is starting up. The point defined by the extrapolation procedure determines the measured next point of the structure.

The black elements within the search matrix are copied into one of two matrices, C1 or C2. First, however, a check is made to determine whether the black elements are already contained in the respective C matrix and a check is also made to investigate whether there is some crossing structure near the search matrix. In the latter case, two points of the crossing structure are stored to give information for its later reconstruction. After that the black elements within the search matrix are copied into the C matrix, and the system goes back to the extrapolation procedure.

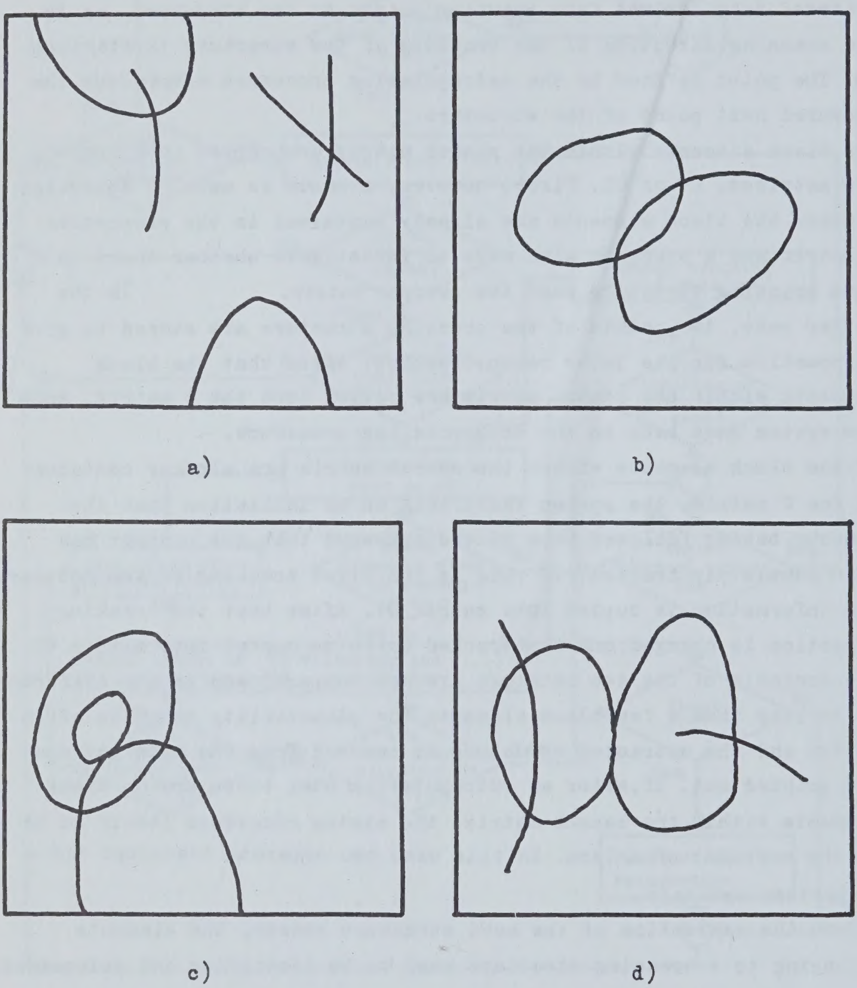
If the black elements within the search matrix are already contained in the C matrix, the system takes this as an indikation that the contour beeing followed is a closed curveand that the contour has been completely tracked. If this is the first tracking of the contour, the information is copied into matrix C1. After that the tracking direction is changed and the tracked curve is copied into matrix C2. The contents of the two matrices are now compared and if the difference is less than a few black elements the plausibility check is affirmative and the extracted structure is remowed from the main pattern and printed out. If,after an extrapolation step there are no black elements within the search matrix, the system considers itself to be at the endpoint of a line. In this case two separate trackings and a comparison are made.

Before the extraction of the next structure starts, the elements belonging to a crossing structure need to be identified and reinserted.

C. Results and conclusions

The program system described has been run on several stylized patterns, where different structures were to be isolated from each other. When constructing suitable patterns it is difficult to get patterns which are optimal for testing the usefulness. Thus it is difficult to give any quantitative result of this pattern processing, as it is always possible to find patterns that it can not process. Instead, some examples and aspects of the system will be discussed.

Figure 5 contains some patterns that have been sucessfully processed and divided up into different structures.



a) b) c) d)

Figure 5. Different patterns processed.

The information for checking on the third level is obtained from the synthesis procedure itself where the checking information is derived in a way different from the other cases. From the printouts it is apparent that this check sorts out a great number of presumptive errors. Other information for the check could also be included if a categorization is made and the reasonability of the derived category is determined. When an error is found in the check, the pattern is scanned from another side and other structures of the pattern become available to the system. In this way the system tries to attack a pattern step by step, isolating structures from parts of the picture where that can easily be done. As the remaining part becomes simpler, the probability of a total partition of the pattern increases. A method which allows the degree of complexity of the pattern to determine the steps of analysis is normally difficult to realize but it can be implemented with trial-and-error procedures.

When the system fails to isolate a structure for several scan periods the tracking is biased in order to increase the probability of isolation of a structure, however at the expense of a higher probability of errors.

The structures representing cell contours obtained from the processed patterns can later be classified with a suitable program. In this case the program used was a program originally used for classification of letters devised by the author [3] .

IV. ACKNOWLEDGEMENTS

The author wishes to thank Professors E. F. Bolinder and B. Wedelin for their valuable suggestions and interest shown during the work.

V. REFERENCES

[1] Wald, A. "Sequential analysis", New York, Wiley, 1947.
[2] Fu, K. S. "Sequential methods in pattern recognition an machine learning, Academic Press, 1968.
[3] Granlund, G. H. "Recognition of letters with Fourier methods", Technical Report Nr 7002, Chalmers Univ. of Technology, Sweden.

SAMPLED-DATA CONTROL SYSTEM DESIGN IN THE PARAMETER PLANE

SHERMAN M. SELTZER
George C. Marshall Space Flight Center, NASA
Huntsville, Alabama

ABSTRACT

This paper describes a graphical technique that permits the sampled-data control system designer to analyze and design a system that will meet desired stability and response specifications. The technique results in a set of contours mapped from the z-domain onto a two parameter plane. The contours display the effect of varying the value of sampling time T, natural frequency ω_n, and their relation to the two chosen adjustable parameters.

INTRODUCTION

The parameter plane technique for analysis and synthesis of linear and nonlinear control systems is amply described in Siljak's recent monograph on the subject.[1] Siljak also has applied the technique to the analysis and synthesis of linear sampled-data control systems.[2] Once the system characteristic equation has been obtained, the parameter plane method enables the designer to evaluate graphically the roots of the equation. Hence, he may design the control system in terms of the chosen performance criteria; e. g., absolute stability, damping ratio, and settling time. He is able to see the effect on the characteristic equation roots of changing two adjustable parameters. Siljak further simplified the design procedure by introducing Chebyshev functions into the equations, thereby putting them in a form particularly amenable to their solution by a digital computer.

In this paper, Siljak's method has been extended to portray the effect of varying the sampling time on the two parameter plane. Thus one can see the effect of the choice of values assigned to the sampling time on absolute and relative stability. Also, the recursive formulas shown herein are simpler in form than the Chebyshev functions.[2] The resulting formulas are cast in a manner that makes them particularly amenable to solution by a digital computer or a calculator. Because a calculator is usually readily available to the design engineer, the examples portrayed here were solved using a Hewlett-Packard desk calculator.

DEVELOPMENT OF THE METHOD

The technique requires that the control system be described by a characteristic equation which is transformed into the z-domain. Two adjustable parameters (k_0, k_1) are selected, and the characteristic equation (CE) is recast in terms of them; i. e.,

$$CE = \sum_{j=0}^{n} \gamma_j z^j = 0 \ , \tag{1}$$

$$\gamma_j = a_j k_0 + b_j k_1 + c_j \ , \tag{2}$$

$$z = \epsilon^{Ts} = r\epsilon^{i\theta} \ , \tag{3}$$

$$r = r(\zeta, \omega_n, T) = \epsilon^{-\zeta \omega_n T} \ , \tag{4}$$

$$\beta = \beta(\zeta, \omega_n, T) = \cos\theta = \cos(\omega_n T \sqrt{1-\zeta^2}) \tag{5}$$

where ζ, ω_n, and T represent the damping ratio, natural frequency, and sampling period, respectively. To transform the characteristic equation from a differential equation into an algebraic equation, z^j may be defined:

$$z^j = X_j + iY_j\sqrt{1 - \beta^2} \tag{6}$$

where $X_0 = 1$, $Y_0 = 0$, $X_1 = r\beta$, and $Y_1 = r$. The following recursion formulas are found to enable one to solve for other values of X_j and Y_j:

$$\begin{aligned} X_{j+1} - 2r\beta X_j + r^2 X_{j-1} &= 0 \\ Y_{j+1} - 2r\beta Y_j + r^2 Y_{j-1} &= 0 \end{aligned} \tag{7}$$

If equation (6) is substituted into equation (1) and the real and imaginary parts of the resulting equation are separated, two simultaneous algebraic equations are obtained. They may be solved for the adjustable gains k_0 and k_1:

$$\begin{aligned} k_0 &= \frac{B_1C_2 - B_2C_1}{J} \\ k_1 &= \frac{A_2C_1 - A_1C_2}{J} \end{aligned} \tag{8}$$

$$J = A_1B_2 - A_2B_1 \tag{9}$$

$$\begin{aligned} A_1 &= \sum_{j=0}^{n} a_jX_j & B_1 &= \sum_{j=0}^{n} b_jX_j & C_1 &= \sum_{j=0}^{n} c_jX_j \\ A_2 &= \sum_{j=0}^{n} a_jY_j & B_2 &= \sum_{j=0}^{n} b_jY_j & C_2 &= \sum_{j=0}^{n} c_jY_j \end{aligned} \tag{10}$$

Recalling that X_j and Y_j are functions of ζ, ω_n, and T and observing that a_j, b_j, and c_j may be functions of T, we see that contours for specified values of ζ may be plotted as functions of ω_n and T in the $k_0 - k_1$ parameter plane. The locations of these contours in the parameter plane represent values of the complex conjugate roots of the characteristic equation. If one chooses numerical values for ω_n and ζ, values for parameters k_0 and k_1 and the sampling period T are established. If the coefficients a_j, b_j, and c_j contain exponential terms with T appearing in the exponents, then each exponent must be replaced by its power series and truncated according to the accuracy that is desired.

The real root locations corresponding to values of z when θ equals 0° and 180° may also be plotted on the parameter plane by setting z equal to a positive or negative real constant, substituting that value into equation (1), and solving for k_1 as a function of k_0. Each resulting contour corresponds to a location of a real root in the z-domain. When $z = +a$,

$$\sum_{j=0}^{n} \gamma_j a^j = 0\ , \tag{11}$$

and when $z = -a$,

$$\sum_{j=0}^{n/2} \omega_{2j}\, a^j - \sum_{j=0}^{n/2-1} \gamma_{2j}\, a^j = 0\ , \tag{12}$$

where a is a positive real number.

STABILITY AND DESIGN

For a linear sampled data control system to be stable, it is necessary that all roots of the characteristic equation lie within the unit circle on the z-plane. It is assumed that the system being designed possesses low-pass filter characteristics so that only the primary strip (corresponding to $0 \leq \theta \leq \pi$ in the z-plane) need be considered. The stable region is bounded by the semi-circle defined by the upper half of the unit circle. This region may be defined by mapping three contours from the z-plane onto the parameter plane. The first contour is defined by setting $\zeta = 0$ in equations (8). The resulting complex root stability boundary is seen to be a function of ω_n and T. The second and third contours correspond to real root locations at $z = +1$ and $z = -1$ and are found by setting "a" equal to unity in equations (11) and (12), respectively. The stable region is determined by applying a shading criterion[2] or using a test point. If the Jacobian, defined by equation (9), is greater than zero, then the stable region (if it exists) lies to the left of the $\zeta = 0$ contour as $\omega_n T$ increases; the left side of the line is double cross-hatched to indicate a boundary associated with double, or complex conjugate, roots. (If the Jacobian is less than zero, the stable region lies to the right.) Single cross-hatching is used on the two contours associated with the real roots. The side of the contour on which to place the cross-hatching is determined by the requirement that cross-hatching be continuous, or on the same side, of the contours as the intersections corresponding to $z = +1$ and $z = -1$ are approached along either the complex root or real root stability boundary.

Once the stable region, if it exists, has been delineated in the parameter plane, the transient characteristics of the system can be determined in terms of the locations of the roots of the characteristic equation. For the complex conjugate roots, these locations are defined in terms of damping ratio and natural frequency. Contours of constant ζ are determined as functions of ω_n and T from equation (8) and plotted on the parameter plane. Similarly, values of "a" corresponding to real root locations may be substituted into equations (11) and (12) to plot contours corresponding to these locations. Now the effect of any chosen design point on the parameter plane may be found in terms of ζ, ω_n, a, and T.

Sometimes it is specified that the settling time of the system be less than a prescribed value. This corresponds to requiring that the real part of the roots of the characteristic equation be less than a prescribed negative real constant. A boundary corresponding to this requirement can be drawn on the parameter plane by mapping a circle of constant radius (for a chosen value of $\omega_n \zeta$ and T) from the z-plane onto the parameter plane.[2] Relations exist for estimating the maximum overshoot and peak time of transient response when it is valid to assume a second order system.[3] However, a simple estimate can sometimes be made by merely looking at the difference equation representing the system response, estimating when the overshoot will occur, and plotting a corresponding line on the parameter plane. This procedure is brought out in the example. Steady state response may be found from the open loop transfer function and the assumed forcing functions in the conventional manner.[3]

EXAMPLE

The technique just described was used in the design and simulation of the control system being developed for NASA's first manned orbital space station, Skylab (Fig. 1).[4] An example of its use in the hybrid simulation of this system is shown. It was found that the system could not be made to behave in the desired manner (in terms of ζ and ω_n) unless the sampling time was made very small, which was undesirable from a computer utilization viewpoint. A digital filter was introduced into the system and analyzed in the parameter plane. Its effect on ζ, ω_n, and, most important for this particular design problem, T is portrayed graphically on the parameter plane.

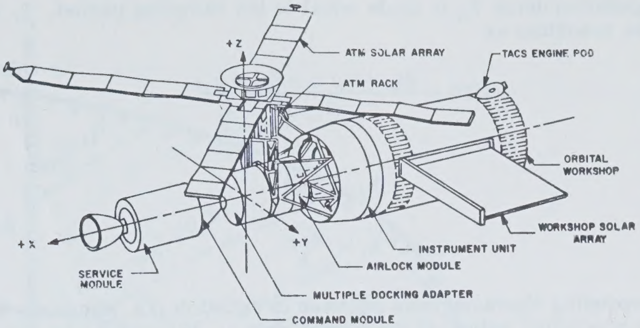

+Z
ATM SOLAR ARRAY
TACS ENGINE POD
ATM RACK
ORBITAL WORKSHOP
WORKSHOP SOLAR ARRAY
+X
+Y
INSTRUMENT UNIT
AIRLOCK MODULE
SERVICE MODULE
MULTIPLE DOCKING ADAPTER
COMMAND MODULE

Figure 1. Skylab

The Skylab control system is shown in block diagram form in Figure 2. If it is assumed that the reaction moment ($\underline{M}_R$) is kept equal to the commanded moment ($\underline{M}_C$) and only rotation about a single axis is considered, the block diagram may be simplified as shown in Figure 3. If at first no digital filter is present, the problem is to choose values for control gains k_0 and k_1 and sampling period T such that desirable transient characteristics accrue. The open loop transfer may be written in the z-domain:

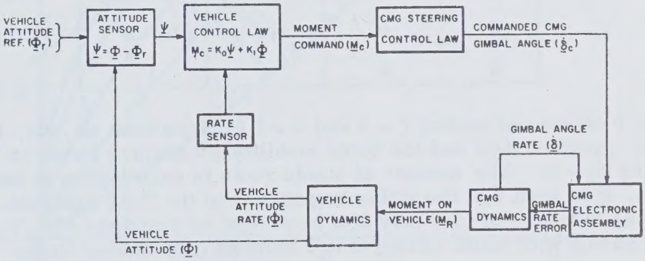

VEHICLE ATTITUDE REF. ($\underline{\Phi}_r$)
ATTITUDE SENSOR
$\underline{\Psi} = \underline{\Phi} - \underline{\Phi}_r$
Ψ
VEHICLE CONTROL LAW
$\underline{M}_c = K_0\underline{\Psi} + K_1\dot{\underline{\Phi}}$
MOMENT COMMAND ($\underline{M}_c$)
CMG STEERING CONTROL LAW
COMMANDED CMG GIMBAL ANGLE ($\dot{\underline{\delta}}_c$)
RATE SENSOR
GIMBAL ANGLE RATE ($\dot{\underline{\delta}}$)
VEHICLE ATTITUDE RATE ($\dot{\underline{\Phi}}$)
VEHICLE DYNAMICS
MOMENT ON VEHICLE ($\underline{M}_R$)
CMG DYNAMICS
GIMBAL RATE ERROR
CMG ELECTRONIC ASSEMBLY
VEHICLE ATTITUDE ($\underline{\Phi}$)

Figure 2. Block Diagram of Skylab Control System

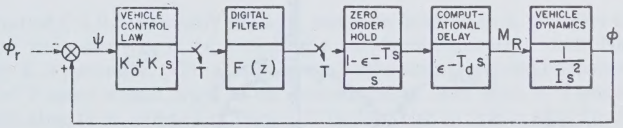

ϕ_r
−
+
ψ
VEHICLE CONTROL LAW
$K_0 + K_1 s$
T
DIGITAL FILTER
F (z)
T
ZERO ORDER HOLD
$\frac{1-\epsilon^{-Ts}}{s}$
COMPUTATIONAL DELAY
$\epsilon^{-T_d s}$
M_R
VEHICLE DYNAMICS
$-\frac{1}{Is^2}$
ϕ

Figure 3. Block Diagram of Simplified Control System

$$\mathcal{Z}G(z) = \left\{\left(\frac{K_0}{s} + K_1\right)\left(\frac{1 - \epsilon^{-Ts}}{s}\right)\left(\epsilon^{-T_d s}\right)\left(\frac{1}{Is^2}\right)\right\} \qquad (13)$$

If the computation delay T_d is made equal to the sampling period, T, then equation (13) may be rewritten as

$$G(z) = \frac{(k_0 + k_1)\, z + (k_0 - k_1)}{z(z - 1)^2} \tag{14}$$

where

$$\begin{aligned} k_0 &= \frac{K_0 T^2}{2I} \\ k_1 &= \frac{K_1 T}{I} \end{aligned} \tag{15}$$

The corresponding characteristic equation is equation (1), where $n = 3$. Table 1 defines the numerical values of the coefficients a_j, b_j, c_j for this case if K_2 is set equal to zero.

Table 1. Coefficients of Characteristic Equation

j	a_j	b_j	c_j
0	1	-1	K_2
1	1	1	$1 - 2K_2$
2	0	0	$K_2 - 2$
3	0	0	1

Using Table 1 values and setting $\zeta = 0$ and $a = 1$ for equations (4), (5), (7), and (8) through (12) provides $J < 0$ and the three stability boundaries shown in Figure 4 (again setting $K_2 = 0$). The number of stable roots in each region is indicated in parentheses. It is seen that the stable region lies in the first quadrant. Contours corresponding to values of $\zeta > 0$ may now be plotted as functions of $\omega_n T$. Several are shown on Figure 5 with equal values of $\omega_n T$ connected by dashed lines.

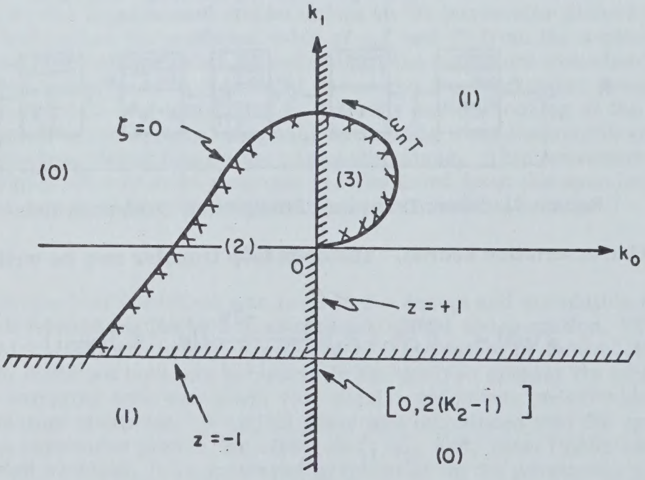


Figure 4. Stability Contours

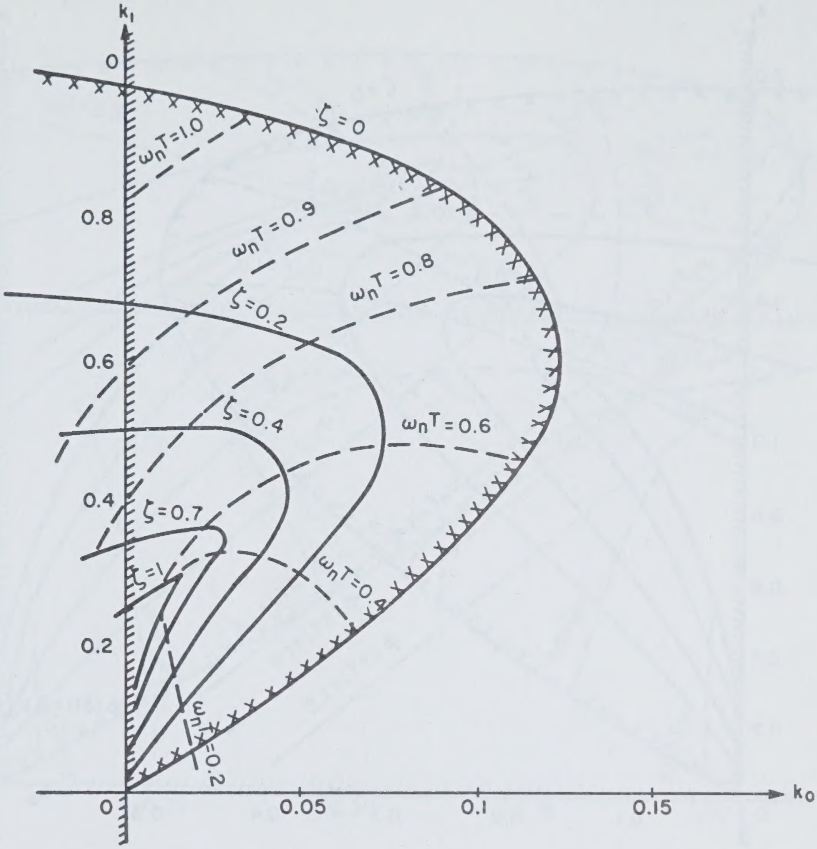


Figure 5. Relative Stability Contours ($K_2 = 0$)

If a value of $\zeta = 0.7$ is desired, $\omega_n T$ must be less than 0.8. The predicted system natural frequency is 0.1π, yielding a maximum possible sampling time of approximately 2.4 seconds. To obtain acceptable transient characteristics, the actual value of T used would have to be considerably less than 2.4 seconds. With the goal of utilizing less computer time, it is desired to add a simple digital filter to increase the sampling period. A filter of the form,

$$F(z) = \frac{z}{z + K_2} , \tag{16}$$

is introduced (Fig. 3). The qualitative effect of the filter may be seen by referring to Figure 4. For positive values of K_2, the $z = -1$ stability boundary is displaced upward. The effect is to move the $\zeta = 0$ curve so that it bulges more to the right and upward, enclosing a larger region and increasing the maximum value of $\omega_n T$ for stability. An example is shown for $K_2 = 1$ (Fig. 6). However, if the value of K_2 is set too large, the $z = -1$ boundary will cut off part (or all) of the stability boundary, as indicated in Figure 7. As K_2 assumes negative values, the $z = -1$ boundary is displaced downward and the $\zeta = 0$ curve moves to the left until, for values of $K_2 < -1$, no stability

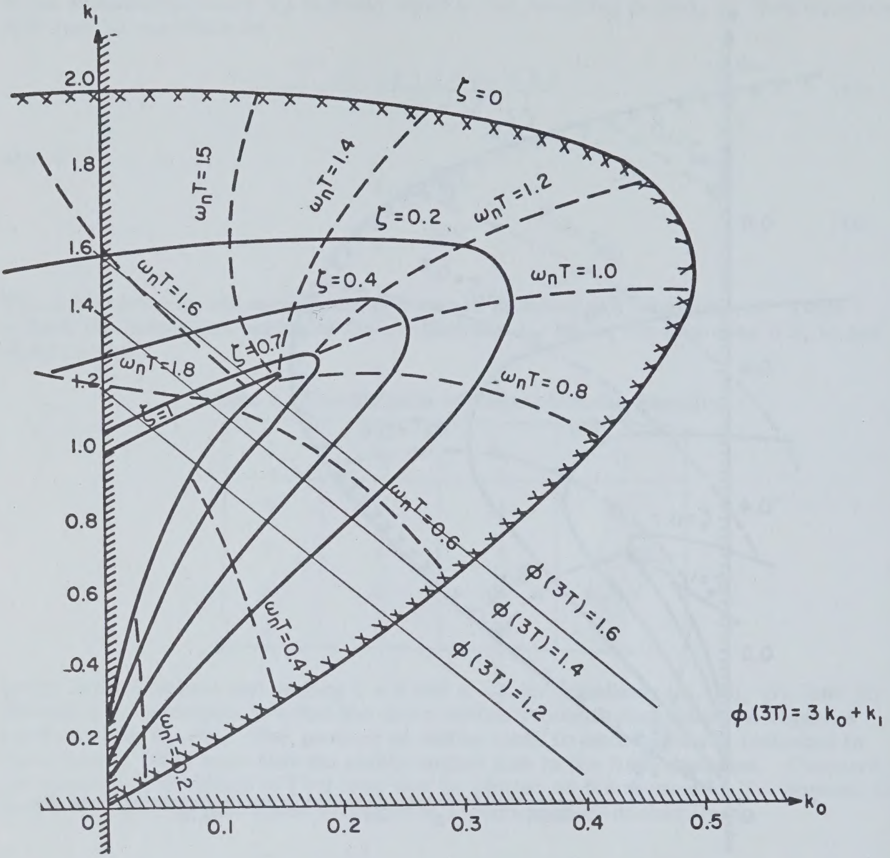


Figure 6. Relative Stability Contours ($K_2 = 1$)

region remains. The steady state error is zero for step displacement and velocity inputs. For acceleration inputs, it is a constant value which is proportional to K_2 and inversely proportional to k_0.

First a value of K_2 should be chosen. It is seen that as K_2 increases positively, larger values of $\omega_n T$ will lie in the stable region for corresponding values of ζ. Although the purpose of the filter is to increase the value of T, K_2 must not be increased to the value where the stable region disappears or to a value greater than the computer clock time. The clock time is 2 5 ms, and the simulation runs are performed at 100 times real time, so the sampling period T must not exceed 2. 5 s ($\omega_n T < 0.79$). Looking at the difference equation associated with the system response ϕ, one sees that in general large values of k_0 and k_1 will result in large overshoots of the response:

$$\begin{aligned} \phi(nT) = {} & (k_0 + k_1)\,\phi_r\,[(n-2)T] + (k_0 - k_1)\,\phi_r\,[(n-3)T] \\ & - (K_2 - 2)\,\phi\,[(n-1)T] - (k_0 + k_1 + 1 - 2K_2)\,\phi\,[(n-2)T] \\ & - (k_0 - k_1 + K_2)\,\phi\,[(n-3)T] \end{aligned} \tag{17}$$

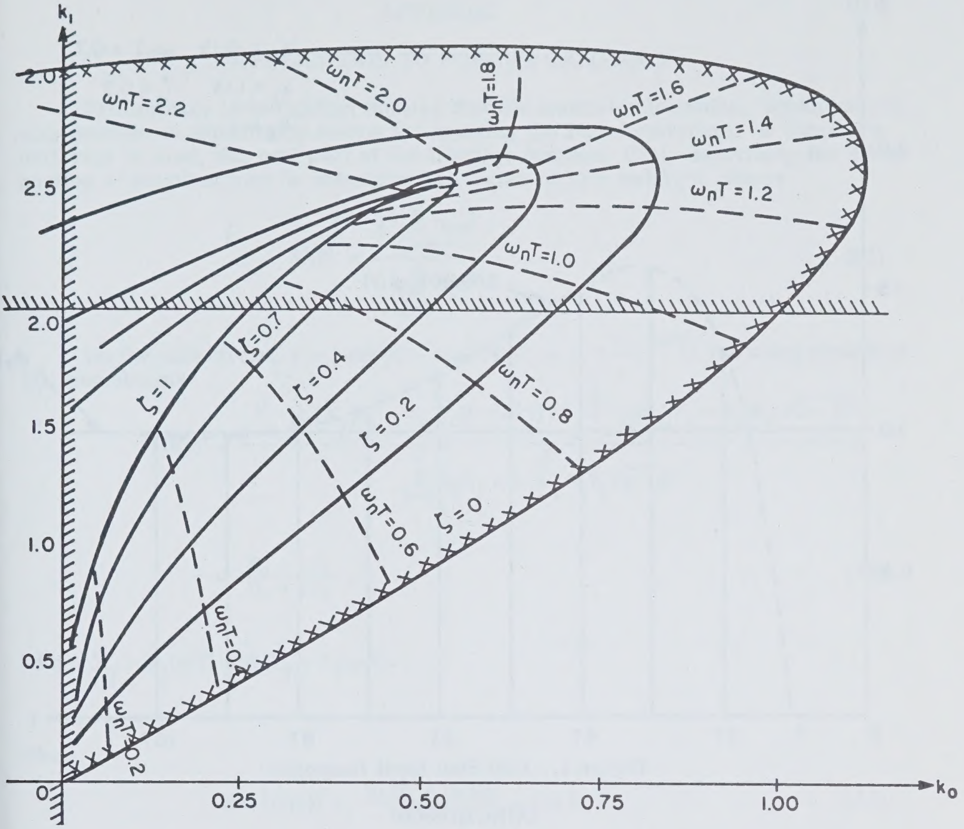


Figure 7. Relative Stability Contours ($K_2 = 2$)

As a design compromise, a value of $K_2 = 1$ was chosen. If a value of 0.707 is used for ζ, the maximum value of $\omega_n T$ that will remain within the stable region is 2.1, which is greater than the clock time constraint. Choosing values for ω_n and ζ has established the locations of the complex conjugate roots; selecting a value for $\omega_n T$ on a particular ζ contour establishes the location of the real root.

If both the difference equation (17) and the parameter plane contours are examined simultaneously, one usually can quickly deduce the sampling period in which the maximum overshoot will occur. For example, if it is assumed that the unit step response will have a maximum overshoot in the vicinity of the third sampling instant, contours corresponding to selected magnitudes of the response at the third sampling instant may be mapped onto the parameter plane, using equation (17). Contours corresponding to unit step response values at the third sampling instant of 1.2, 1.4, and 1.6 are indicated on Figure 6. If it is desired that the overshoot not exceed 60%, then it is estimated that $1.5 < \omega_n T < 0.7$. Because of the clock time constraint, $\omega_n T$ is set at the lower value. The resulting step input response is shown on Figure 8.

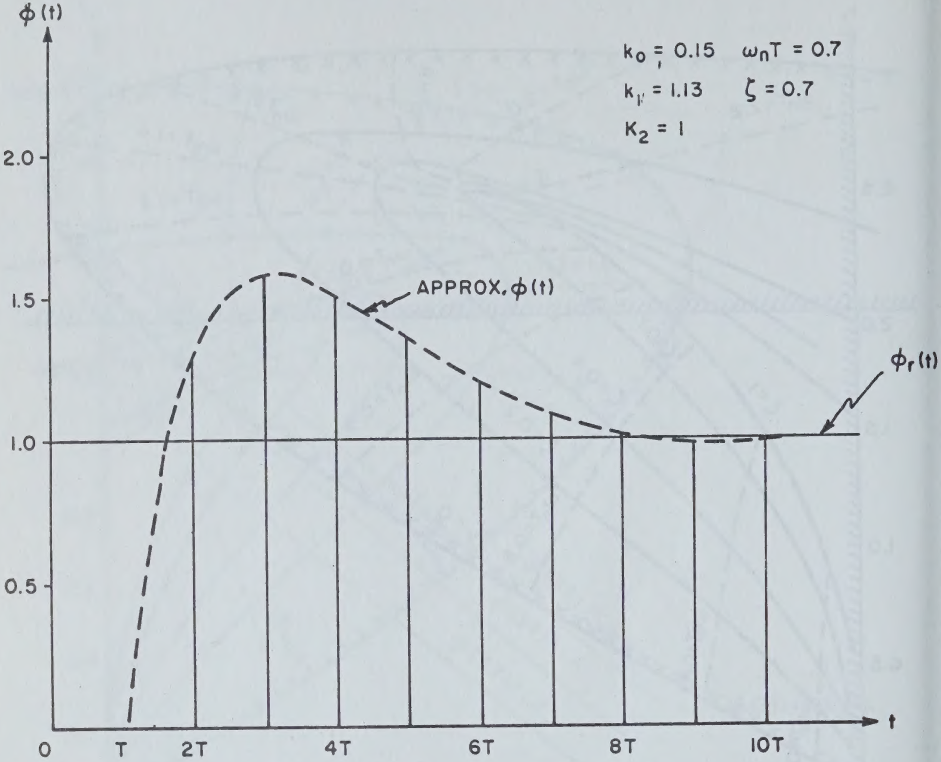


Figure 8. Unit Step Input Response

CONCLUSION

The graphical technique described was applied to a preliminary design problem and found to be easy and rapid to implement. Although the example is simple, the technique should be readily applicable to more complicated problems, particularly if a digital computer is used as a design tool, for the equations have been formulated in a manner that makes them amenable to digital computer solution. It has been shown how the technique may be supplemented by analytically calculating steady state responses and by examining the difference equation describing the system response.

REFERENCES

1. Siljak, D. D., Nonlinear Systems, Wiley, N. Y., 1969.

2. Siljak, D. D., Analysis and Synthesis of Feedback Control Systems in the Parameter Plane, Pt. II - Sampled-Data Systems, IEEE Trans., Pt. II (Applications and Industry), 83, 458-466 (Nov. 1964).

3. Kuo, B., Analysis and Synthesis of Sampled-Data Control Systems, Prentice Hall, N. J., 1963.

4. Seltzer, S. M., and Chubb, W. B., Skylab Attitude and Pointing Control System, NASA TN D-6068, Oct. 1970.

APPENDIX

APPLICATION TO POPOV'S CRITERION

Preliminary investigation implies that the recursion formulas, equations (7), may provide an expeditious means for applying the Popov criterion. If Tsypkin's criterion is used, the real part of the transfer function, G(z), describing the linear portion of the plant may be determined in terms of X_j's and Y_j's, where

$$G(z) = \frac{z^p \sum_{j=0}^{n} \gamma_j z^j}{\sum_{j=0}^{d} \eta_j z^j} \tag{A1}$$

On the unit circle, $r = 1$ and $\beta = \cos \omega T$ (i.e., $z = \epsilon^{i\omega T}$) and using equations (7), one obtains

$$G(z) = \frac{\sum_{j=0}^{n} \gamma_j [X_p X_j - Y_p Y_j (1 - \beta^2)] + i \sum_{j=0}^{n} \gamma_j (X_p Y_j + Y_p X_j) \sqrt{1 - \beta^2}}{\sum_{j=0}^{d} \eta_j X_j + i \sum_{j=0}^{d} \eta_j Y_j \sqrt{1 - \beta^2}}$$

$$= \frac{N_1 + iN_2}{D_1 + iD_2}, \tag{A2}$$

where $X_j = X_j(\omega T)$ and $Y_j = Y_j(\omega T)$.

Then

$$\mathscr{R}e\left\{G(z)\right\} = \frac{N_1 D_1 + N_2 D_2}{D_1^2 + D_2^2} = f(\omega T, T) \tag{A3}$$

Contours representing $\mathscr{R}e\{G(z)\}$ versus ωT may be plotted for selected values of T. Using Tsypkin's criterion,

$$\mathscr{R}e\left\{G(z)\right\}\Big|_{z = \epsilon^{i\omega T}} > -\frac{1}{k} \tag{A4}$$

one may find values of the limiting slope k, for the acceptable nonlinear sector.

INITIAL CONDITION ZONES FOR IPFM SYSTEM STABILITY

Warren J. Guy Jr., Marvin Kurland and Andrew Meyer
Lafayette College Newark College of Engineering
Easton, Pa. Newark, New Jersey

ABSTRACT

Initial condition zones are determined which guarantee system stability for the class of IPFM systems considered. The pulse polarity sequence which drives the plant to the origin is also given.

INTRODUCTION

Pulse frequency modulated control systems, and in particular Integral Pulse Frequency Modulated (IPFM) systems are becoming an increasingly popular field of research. The IPFM modulator functions by emitting an impulse when the integral of the input signal reaches a certain constant threshold value, A. The impulse times are thus unknown a priori and their determination constitutes the difficulty with this class of system. The constant strength pulse, M, which is emitted from the modulator has a positive polarity if the integral of the input reaches +A and a negative polarity if -A is obtained by the integral.

The autonomous system with a single modulator and linear element is very sensitive to the system initial conditions. For a given system with initial conditions in a small region, the resulting motion may produce periodic or aperiodic oscillations or a stable trajectory (motion will go to the origin). In particular it is desired to know what initial conditions will yield a stable motion without the necessity of computing impulse times. By mapping backwards a capture zone about the origin, it is possible to define those regions in which the initial condition response will drive the linear plant to "0".

ZONE DETERMINATION

The technique is similar to a construction by Jury and Blanchard (1) for determining the trajectory from a given set of initial conditions. Define first a zone, $\overline{R(0)}$, about the origin such that,

$$A > \max_{t} \left| \int_0^t e(\tau)\, d\tau \right|$$

$$> \max_{t} \left| \underline{c}^T \underline{A}^{-1} (e^{\underline{A}t} - \underline{I}) \underline{x}_0 \right|$$

where $\underline{A}^{-1}$ will exist for the class of linear elements considered. $\underline{x}_0$ determines the points within the zone and thus the initial conditions which produce no output from the modulator.

Between impulses the system operates as though it were opened loop and thus may be treated as a new problem after each impulse. This gives rise to the following algorithm for the initial condition zone determination which ulimately drives the linear element to the origin in "n" impulses from the modulator:

$$\underline{x}(t_n^-) = \underline{x}_{0(n-1)} - b_n M\underline{b} \qquad \underline{x}_{0(n-1)} \in R(n-1)$$

$$\underline{c}^T\underline{A}^{-1}\{\underline{x}(t_n^-) - \underline{x}_{0n}\} = -b_n A$$

$$\underline{x}_{0n} = e^{-\underline{A}t_n}\,\underline{x}(t_n^-)$$

where $\underline{x}$ is the state of the system, t_n is the impulse time, $R(n-1)$ denotes the region for stability for $(n-1)$ impulses and $\underline{x}_{0n}$ the initial conditions for stability in "n" impulses. See figure 1 below.

The above algorithm can easily be adapted to a construction method on the state plane as the example below illustrates, figure 2.

CONCLUSION

This work is part of a program to predict the motion of single and compound IPFM systems given the system initial conditions. Support for the research has been given in part by the National Science Foundation.

REFERENCES

1. Jury, E.I. and J. Blanchard, "A Nonlinear Discrete Equivalence of Integral Pulse Frequency Modulated Systems," IEEE Trans AC, Vol. 12, August 1967, pp. 415-422.

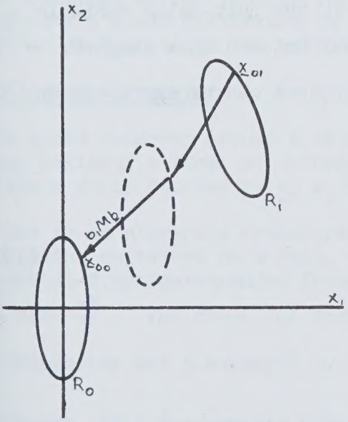


Figure 1. Backward mapping

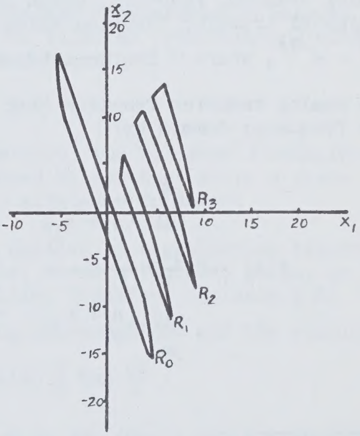


Figure 2. Example $\underline{A} = \begin{vmatrix} 0 & 1 \\ -2 & -3 \end{vmatrix}$

$\underline{c}=(1,0)^T$, $\underline{b}=(0,1)^T$

Z TRANSFORMS AS HYPERBOLIC TANGENT TRANSFORMATIONS

STANLEY A. WHITE
Autonetics Division, North American Rockwell Corp., Anaheim, Calif.

I. SUMMARY

Digital filter design is frequently based on digitization of the transfer function of a continuous filter, the analog prototype. The digitization of the analog design is executed by applying the appropriate z-transformation to the continuous-filter transfer function. All-pole lowpass designs can use the impulse-invariant (standard) transformation. Bandstop or flat bandpass characteristics are preserved best by a second transform, the bilinear (or Tustin) transformation, but difficulty is encountered in controlling slopes or phase shift which contribute to transient-response distortion. Critical frequencies must be pre-warped, however, because of frequency-plane distortion inherent in the transformation. An easy-to-apply transformation that reasonably preserves both the transient and frequency response characteristics is a third transform, the matched z-transformation. Filters digitized by the matched z-transformation exhibit many of the favorable characteristics of both the impulse-invariant and bilinear-transform-derived designs.

This paper presents a unified viewpoint by which each of these transformations is viewed as an incompletely-performed hyperbolic-tangent transformation. The natural consequence of this is the suggestion of new transformations which are simply more-completely-performed hyperbolic-tangent transformations.

II. REVIEW OF CONVENTIONAL Z-TRANSFORMS*

The z-transforms that we are considering are a means of defining digital approximations to analog, or continuous, transfer functions. The primary objective of a z-transformation can be viewed as the conversion of an analog transfer function, which is rational in the Laplace variable, s, to a digital transfer function which is rational in the unit-delay variable $z^{-1} = e^{-Ts}$, where T is the uniform time interval between data samples.

The analog transfer function that is to be digitized can be expressed in the frequency domain by:

$$F(s) = K \frac{\prod_{i=1}^{m} (s + a_i)}{\prod_{j=1}^{n} (s + b_j)} = \sum_{j=1}^{n} \frac{K_j}{s + b_j} \tag{1}$$

---

*Various time domain interpretations of various z transformations are acknowledged, but not pursued here as germaine to the discussion (Ref. 1-3).

where $-a_i$ and $-b_j$ are the generally complex zeros and poles of the transfer function (the b_j's are assumed to be unrepeated if the partial fraction expansion form is used), the K_j are the residues of the poles, and $m < n$.

Impulse-Invariant or Standard z-Transform

In order to apply the standard, or impulse-invariant z-transform, the transfer function of the continuous filter is expressed as a weighted (by residues) sum of transfer functions, which contain first-order poles only, and which was obtained by partial-fraction expansion.

The sampled-data transfer function is then obtained by replacing each $(s + a)$ term by $1 - e^{-aT} z^{-1}$, where T is the sampling period and $z^{-1} = e^{-sT}$ is the unit-delay operator.

$$F(z) = \sum_{j=1}^{n} \frac{K_j}{1 - e^{-b_j T} z^{-1}}$$

Note that $\lim_{T\to 0} F(z) = \lim_{T\to 0} \frac{1}{T} F(s)$. Because the digital filter frequency is followed by a D/A converter which usually contains a zero-order hold,

$$H_0(s) = \frac{1 - e^{-Ts}}{s}, \text{ and consequently } \lim_{T\to 0} F(z)\, H_0(s) = F(s).$$

This transformation exhibits some serious deficiencies, particularly when the preservation of frequency responses of other than lowpass functions is of prime importance.

Fortunately, the design of bandstop and highpass digital filters can be accomplished by using either the bilinear z-transform or the matched z-transform by sacrificing the transient response characteristic because of severe phase distortion in the case of the bilinear transform and mild phase distortion in the case of the matched z-transform.

Bilinear z-Transform

To avoid foldover errors a second transformation, the bilinear z-transform (or bilinear z-form, or z-form), was developed to map the entire s-plane into a strip bounded by $\pm j\,\omega_s/2$ in order to eliminate foldover.

This is an algebraic transform that can be applied to a continuous transfer function expressed as a ratio of polynomials, or as a factored ratio, or in partial-fraction-expanded form. The digitizing transform replaces s by $\frac{2}{T} \tanh \frac{sT}{2}$. The resultant frequency warping is recognized and the critical frequencies are pre-warped by replacing ω with $\frac{2}{T} \tan \frac{\omega T}{2}$.

However, this pre-warping usually is applied to the imaginary part only of the critical frequencies. That is, as the transform exists, each factor $s + a$ is replaced by:

$$\frac{2}{T}\tanh\frac{sT}{2} + \mathrm{Re}(a) + \frac{2}{T}\tanh\left\{\frac{jT}{2}\,\mathrm{Im}(a)\right\} .$$

Through the mechanics of this important transformation, the Laplace variable s is replaced by the digitized Laplace variable s_1, where their relationship is described by:

$$\left[\, s = \frac{2}{T}\tanh\frac{s_1T}{2} = \frac{2}{T}\,\frac{e^{\frac{s_1T}{2}} - e^{-\frac{s_1T}{2}}}{e^{\frac{s_1T}{2}} + e^{-\frac{s_1T}{2}}} = \frac{2}{T}\,\frac{1 - z^{-1}}{1 + z^{-1}} \,\right]$$

This implies that the real and imaginary components of the roots of the transfer function are similarly distorted:

$$\sigma + j\omega = \frac{2}{T}\tanh\frac{(\sigma_1 + j\omega_1)}{T^2}\;T$$

Separating the real and imaginary parts yields the following pair of relationships:

$$\omega = \frac{2}{T}\tan\frac{\omega_1T}{2}\;\frac{1 - \tanh^2\frac{\sigma_1T}{2}}{1 + \tan^2\frac{\omega_1T}{2}\tanh^2\frac{\sigma_1T}{2}}$$

$$\sigma = \frac{2}{T}\tan\frac{\sigma_1T}{2}\;\frac{1 + \tan^2\frac{\omega_1T}{2}}{1 + \tan^2\frac{\omega_1T}{2}\tanh^2\frac{\sigma_1T}{2}}$$

This says that both the imaginary and real parts of the roots should be pre-warped. However, as stated above, usually the only pre-warping done is on the imaginary part of the complex roots so that:

$$\omega = \frac{2}{T}\tan\frac{\omega_1T}{2}$$

and

$$\sigma = \sigma_1$$

which are correct results only for very small values of σ. However, to obtain the sampled-data bilinear z-transform set

$$F(z) = F_1(s) \Big|_{s = \frac{2}{T}\frac{1 - z^{-1}}{1 + z^{-1}}}$$

where $F_1(s)$ is the pre-warped form of $F(s)$. This transform yields designs that have excellent bandpass or bandstop characteristics, but with poor phase and slope control, and inferior impulse response to that of the impulse-invariant z-transform.

Matched-z Transform

A third transformation that enjoys wide usage and does not require that a partial-fraction expansion be made is the so-called "matched-z" transform. The matched-z transform of the equation is usually expressed as:

$$F(z) = K_1 \frac{\prod_{i=1}^{m} \left(1 - z^{-1} \quad e^{-b_j T}\right)}{\prod_{j=1}^{n} \left(1 - z^{-1} \quad e^{-b_j T}\right)}$$

This transform is suggested by the periodicity of the standard z-transform; its outstanding credentials are that it is very simple to obtain, and that it works well, yielding digital filter designs capable of giving "good" high-frequency (near the Nyquist rate) performance as compared to those digital filters obtained by use of the standard z-transform.*

Note that the poles of the filter function are the same as those derived by the standard z-transform. However, the zeros of the function will not usually correspond to those for the standard z-transform. The results of this fact is that the matched z-transform may be used to obtain useful designs for highpass and bandstop filters, but contributes to the distortion. There are some designs in which the filter response is unsatisfactory, unless some additional zeros are added at half the sampling frequency. Then the appearance of the matched z-transform becomes:

$$F(z) = K_1 \frac{\prod_{i=1}^{m} \left(1 - z^{-1} \quad e^{-a_i T}\right)}{\prod_{j=1}^{n} \left(1 - z^{-1} \quad e^{-b_j T}\right)} \quad (1 + z^{-1})^N$$

where N is empirically determined.

---

*Kaiser suggests that "high frequency" in a sampled-data system be taken to mean on the order of 70 to 80 percent of the Nyquest frequency.

An Observation

The substitution used in defining the bilinear z-transform

$$s \Rightarrow \frac{2}{T}\,\frac{1 - z^{-1}}{1 + z^{-1}} = \frac{2}{T}\tanh\frac{sT}{2}$$

has the pleasant property of being rational in z^{-1}

If we replaced s by $\frac{2}{T}\tanh(Ts/2)$ and b by $\frac{2}{T}\tanh(bT/2)$ in (1), we would have:

$$s + b \Rightarrow \frac{2}{T}\,\frac{1 - z^{-1}}{1 + z^{-1}} + \frac{2}{T}\,\frac{1 - e^{-bT}}{1 + e^{-bT}} \quad \frac{4}{T\,(1 + e^{-bT})}\,\frac{1 - e^{-bT}\,z^{-1}}{1 + z^{-1}}$$

If this substitution pair is made on the partial-fraction expanded form of F(s), the following results:

$$F(s) = \left.\vphantom{\sum_{j=1}^{n}\frac{T K_j^1}{1 - e^{-b_jT} z^{-1}}}\right|_{\substack{s \,=\, \frac{2}{T}\tanh\frac{sT}{2} \\ b_j \,=\, \frac{2}{T}\tanh\frac{b_jT}{2}}} = (1 + z^{-1}) \sum_{j=1}^{n} \frac{T\,K_j^1}{1 - e^{-b_jT}\,z^{-1}}$$

where $K_j^1 = K_j\left(1 + e^{-b_jT}\right)/4$

which bears a striking resemblance to the impulse-invariant z-transform. If this substitution pair is made on the factored form of F(s), the following results:

$$F(s) \left.\vphantom{\frac{\prod_{k=1}^{m}}{\prod_{j=1}^{n}}}\right|_{\substack{s \,=\, \frac{2}{T}\tanh\frac{sT}{2} \\ a \,=\, \frac{2}{T}\tanh\frac{aT}{2} \\ b \,=\, \frac{2}{T}\tanh\frac{bT}{2}}} = K_2\,\frac{\prod\limits_{k=1}^{m}\left(1 - e^{-a_kT}\,z^{-1}\right)}{\prod\limits_{j=1}^{n}\left(1 - e^{-b_jT}\,z^{-1}\right)}\,(1 + z^{-1})^{M}$$

where

$$K_2 = \left(\frac{4}{T}\right)^{m-n} \frac{\prod_{j=1}^{n} \left(1 + e^{-b_j T}\right)}{\prod_{k=1}^{m} \left(1 + e^{-a_k T}\right)}$$

and $M = n - m$

which is the result (within an amplitude-scaling coefficient) of the result obtained by the conventional matched z-transform.

From this it can be seen that complete complex pole pre-warping yields the same pole locations as the impulse-invariant z-transformation. The zeros are different (and easier to obtain) and the additional zeros that appear at half the sampling frequency appear rather naturally, rather than by magic.

III. THE HYPERBOLIC-TANGENT TRANSFORMATION

Another Observation

Rather than using the piecemeal substitutions discussed above, a logical next step is to digitize transfer functions by making a hyperbolic-tangent transformation on entire factors. Since

$$\tanh (x + y) = \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y}$$

then

$$s + b = \frac{2}{T} \tanh \frac{T}{2} (s + b) = \frac{2}{T} \frac{1 - e^{-bT} z^{-1}}{1 + e^{-bT} z^{-1}}$$

If this warping is treated as a digitizing transformation and applied to the partial-fraction expanded transfer function, the result is

$$F(s) \Bigg|_{s + b = \frac{2}{T} \tanh \frac{T}{2} (s + b)} = \frac{T}{2} \sum_{j=1}^{n} K_j \frac{\left(1 + e^{-b_j T} z^{-1}\right)}{\left(1 - e^{-b_j T} z^{-1}\right)}$$

If this is applied to the factored transfer function, the result is

$$F(s)\Bigg|_{\substack{s + a = \frac{2}{T} \tanh \frac{T}{2} (s + a) \\ s + b = \frac{2}{T} \tanh \frac{T}{2} (s + b)}}$$

$$= \left(\frac{T}{2} \right)^{n-m} K \frac{\prod_{i=1}^{m} \left(1 - e^{-a_i T} z^{-1}\right) \prod_{j=1}^{n} \left(1 + e^{-b_j T} z^{-1}\right)}{\prod_{j=1}^{n} \left(1 - e^{-b_j T} z^{-1}\right) \prod_{i=1}^{m} \left(1 + e^{-a_i T} z^{-1}\right)}$$

The curious geometrical interpretation of this in the z plane is that there is a cluster of poles and zeros added in the z plane, a cluster that is the mirror image (reflected across the origin of the z plane) of the original pole-zero set obtained by the matched-z transform (except for the zeros to be added at the Nyquist frequency), with the left-half plane poles and zeros interchanged.

At low frequencies, the mirror image pole-zero cluster looks very nearly like that required cluster of zeros at the Nyquist frequency that arises when the matched-z transform is used. At higher frequencies (above half-Nyquist frequency), this mirror-image cluster sometimes shapes the amplitude and phase responses of this digital filter to behave more like those of its analog prototype than those of conventional digital filters.

IV. CONCLUSION

It is premature to draw a blanket conclusion, however, when various digital-filter structures and transformations from analog prototypes are being evaluated as the basis for design of digital filters for specific applications, the hyperbolic-tangent-transform-derived filter should be investigated as a serious candidate. More research should be conducted in this area to define the types of applications for which this transformation is superior to others.

V. REFERENCES

1. H. R. Martens, "A Comparative Study of Digital Integration Methods," Simulation, pp 87-94, February 1969.

2. S. S. Haykin and R. Carnegie, "New Method of Synthesizing Linear Digital Filters Based on Convolution Integral," Proc. IEE, Vol. 117, No. 6, pp 1063-1071, June 1970.

3. C. M. Rader, "On Digital Filtering," IEEE Trans., AU-16, No. 3, pp 303-314, September 1968.

STABILIZATION OF A CLASS OF NONLINEAR SAMPLED-DATA SYSTEMS USING A LINEAR SAMPLED-DATA SYSTEM ESTIMATOR

MOHAMED I. YOUNIS WILLIAM G. VOGT
Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, Pa. 15213

ABSTRACT

It has been shown that the state of a linear system can be constructed from observations of its inputs and outputs. The observer which performs the reconstruction is itself a linear system with arbitrary time constants. This linear asymptotic estimator has been used previously to stabilize certain types of continuous nonlinear systems. In this paper, it is shown that a linear sampled-data asymptotic estimator can be used to construct the state of a linear sampled-data system, and this will be used to stabilize a class of nonlinear sampled-data systems.

INTRODUCTION

Many designs of modern control theory are based on the assumption that the state vector of the system to be controlled is available for measurements. However, in many practical applications, only a few output quantities are available for measurements. Luenberger (1) showed that the state vector of a linear system can be constructed from observations of its inputs and outputs. The observer which performs the reconstruction is itself a linear system with arbitrary time constants. Strane and Vogt (2) showed that for a class of nonlinear continuous systems an estimator, in the sense of Luenberger, can be designed to ensure the stability of the system. In this paper a sampled-data state estimator in the sense of Luenberger is developed which will be used to stabilize (1) a linear sampled-data system (LSDS), and (2) a Lur'e type nonlinear sampled-data system. The latter is described by the difference equations

$$\underline{x}((k+1)T) = \underline{\Phi}(T)\underline{x}(kT) + \underline{\psi}(T)\underline{b}\theta(\sigma) \tag{1}$$

$$\sigma = \underline{c}'\underline{x}(kT) \tag{2}$$

$$y_1 = \theta(\sigma) \tag{3}$$

$$y_2 = \underline{d}'\underline{x}(kT) \tag{4}$$

where $\underline{x}(kT)$, $\underline{\psi}(T)\underline{b}$, $\underline{c}$ and $\underline{d}$ are n-vectors, y_1 and y_2 are available system outputs, and $\theta(\sigma)$ is a single valued, continuous real function, which is defined for all real values of its argument, σ, and which satisfies the conditions

$$\theta(0) = 0; \quad \dot{\theta}(\sigma) \geq 0; \qquad 0 \leq \frac{\theta(\sigma)}{\sigma} \leq k \quad \text{for } \sigma \neq 0 \tag{5}$$

$$\lim_{\sigma\to\infty} \int_0^\sigma \theta(\sigma)d\sigma = \infty \tag{6}$$

In order to prove the above assertions it is necessary to make use of the following theorems:

Theorem 1: (3) Consider a LSDS described by

$$\underline{x}((k+1)T) = \underline{\Phi}(T)\underline{x}(kT) + \underline{h}(T)u(kT) \tag{7}$$

$$\underline{y}(kT) = \underline{x}(kT) \tag{8}$$

where $(\underline{\Phi},\underline{h})$ is completely controllable. If feedback is applied from the "clean" output, $\underline{y}$, to the control scalar u, by a feedback matrix, k', then, the eigenvalues of the stabilized system transition matrix, $(\underline{\Phi}(T)+\underline{h}(T)k')$

can be arbitrarily chosen by adequate selection of the components of the feedback matrix $\underline{k}'$.

The proof of this theorem follows from the proof of the continuous case by Kalman.(4)

Theorem 2:(3) Consider a LSDS described by

$$\underline{x}((k+1)T) = \underline{\Phi}(T)\underline{x}(kT) + \underline{I}\ \underline{u}(kT) \tag{9}$$

$$y(kT) = \underline{c}'\underline{x}(kT) \tag{10}$$

where $[\underline{\Phi}(T),\underline{c}']$ is completely observable. If feedback is applied from the output, $y(kT) = \underline{c}'\underline{x}(kT)$, to the inputs $u_i(kT)$, $i = 1,2,\ldots,n$, by an arbitrary feedback matrix, $\underline{k}$, then the eigenvalues of the stabilized system transition matrix, $(\underline{\Phi}(T) + \underline{k}\ \underline{c}')$ can be arbitrarily chosen by adequate selection of the feedback matrix, $\underline{k}'$.

The proof of this theorem follows from Luenberger(1) and Strane and Vogt(2).

Theorem 3: (Tsypkin(5) Criterion for stability of non-linear sampled-data systems). Let $p(z)$ and $q(z)$ denote polynomials without common factors and let

$$\underline{x}(k+1) = \underline{A}\ \underline{x}(k) + \underline{b}u(k)\ ;\ y(k) = \underline{c}'\underline{x}(k) \tag{11}$$

be an irreducible representation of $G(z) = \frac{q(z)}{p(z)}$. If the zeros of $p(z)$ lie inside the disk $|z| = 1$, then the null solution of

$$\underline{x}(k+1) = \underline{A}\ \underline{x}(k) - \underline{b}f(\underline{c}'\underline{x}(k)) \tag{12}$$

is asymptotically stable in the large if

$$0 \leq \frac{\theta(\sigma)}{\sigma} \leq k,\ \underline{\theta}'(\sigma) \geq 0 \tag{13}$$

and $$\mathrm{Re}[1+\alpha(1-z^{-1})G(z)] + \frac{1}{k} > 0 \quad \text{for } |z| = 1 \tag{14}$$

LINEAR SAMPLED-DATA SYSTEM COMPENSATION USING SAMPLED-DATA STATE ESTIMATOR

The Linear Sampled-Data System

The LSDS to be considered is shown by the signal flow graph given in figure (1) to the left of the center line. The difference equation which describes the system is given by

$$\underline{x}((k+1)T) = \underline{\Phi}(T)\underline{x}(kT) + \underline{\psi}(T)\underline{b}u(kT) \tag{15}$$

where

$$\underline{\Phi}(T) = e^{\underline{A}T},\ \underline{\psi}(T) = \int_0^T e^{A(T-\tau)}\,d\tau \tag{16}$$

The Sampled-Data State Estimator

Luenberger(1) has presented a state estimator for the continuous system. It is in this sense that a sampled-data estimator is presented here. This estimator is shown by the signal flow graph given in figure (1) to the right of the center line. After some manipulation the difference equation which describes the estimator is given by

$$\hat{\underline{x}}((k+1)T) - \underline{x}((k+1)T) = [\underline{\Phi}(T) - \underline{\psi}(T)\underline{g}\ \underline{c}'][\hat{\underline{x}}(kT) - \underline{x}(kT)] \tag{17}$$

where $\hat{\underline{x}}(kT)$ represents the estimated value of the state vector $\underline{x}(kT)$.

The solution to this vector matrix difference equation is

$$\hat{\underline{x}}(kT) - \underline{x}(kT) = [\underline{\Phi}(T) - \underline{\psi}(T)\underline{g}\ \underline{c}']^k(\hat{\underline{x}}(o) - \underline{x}(o)) \tag{18}$$

or

$$\begin{aligned}\hat{\underline{x}}(kT) &= \underline{x}(kT) + [\underline{\Phi}(T) - \underline{\psi}(T)\underline{g}\ \underline{c}']^k(\hat{\underline{x}}(0) - \underline{x}(0))\\ &= \underline{x}(kT) + [\underline{\Phi}(T) - \underline{\psi}(T)\underline{g}\ \underline{c}']^k\ \tilde{\underline{x}}(0)\end{aligned} \tag{19}$$

where $\tilde{\underline{x}}(0) = \hat{\underline{x}}(0) - \underline{x}(0)$ is the error in the estimate of $\underline{x}(kT)$.

From theorem 2, the eigenvalues of $[\underline{\Phi}(T) - \underline{\psi}(T)\underline{g}\ \underline{c}']$ can be arbitrarily specified, e.g., less than one in magnitude, by proper choice of $\underline{g}$, which is completely arbitrary, provided only that $[\underline{\Phi}(T),\underline{c}']$ is completely observable. Thus, the coefficient of the error, $\tilde{\underline{x}}$, in (6) approaches zero as k increases, and $\hat{\underline{x}}(kT) \to \underline{x}(kT)$ as $\underline{k} \to \infty$. This means that for a completely observable LSDS described by (3-4) a sampled-data asymptotic estimator can always be constructed such that the estimated vector $\hat{\underline{x}}(kT)$ is equal to the state vector $\underline{x}(kT)$ after the error dies out.

Thus, the following theorem can be stated.

Theorem 4: Given a completely observable LSDS

$$\underline{x}((k+1)T) = \underline{\Phi}(T)\underline{x}(kT) + \underline{\psi}(T)\underline{b}u(kT) \tag{20}$$

$$y(kT) = \underline{c}'x(kT) \tag{21}$$

a sampled-data symptotic estimator can always be constructed such that the estimated vector $\hat{\underline{x}}(kT)$ is equal to the system state vector $\underline{x}(kT)$ plus an error $[\underline{\Phi}(T) - \underline{\psi}(T)\underline{g}\ \underline{c}']^k(\hat{\underline{x}}(o) - \underline{x}(o))$. Furthermore, the dynamics of the error term $[\underline{\Phi}(T) - \underline{\psi}(T)\underline{g}\ \underline{c}']^k(\tilde{\underline{x}}(o))$ are completely arbitrary.

System Compensation Using the State Estimator

Consider figure (1) in which a sampled-data asymptotic estimator is cascaded to a LSDS with feedback indicated by the dotted line. From the presentation above, the vector matrix difference equations which describe the system plus the estimator may be given by

$$\underline{x}((k+1)T) = \underline{\Phi}(T)\underline{x}(kT) + \underline{\psi}(T)\underline{b}\ \underline{k}'\hat{\underline{x}}(kT) + \underline{\psi}(T)\underline{b}u(kT) \tag{22}$$

$$\begin{aligned}\hat{\underline{x}}((k+1)T) &= \underline{\Phi}(T)\hat{\underline{x}}(kT) + \underline{\psi}(t)\underline{b}\ \underline{k}'\hat{\underline{x}}(kT) + \underline{\psi}(T)\underline{b}u(kT)\\ &\quad + \underline{\psi}(T)\underline{g}\ \underline{c}'\underline{x}(kT) - \underline{\psi}(T)\underline{g}\ \underline{c}'\hat{\underline{x}}(kT)\end{aligned} \tag{23}$$

Equations (22) and (23) can be expressed as a set of 2n difference equations

$$\begin{bmatrix}\underline{x}((k+1)T)\\ \hat{\underline{x}}((k+1)T)\end{bmatrix} = \begin{bmatrix}\underline{\Phi}(T) & \underline{\psi}(T)\underline{b}\ \underline{k}'\\ \underline{\psi}(T)\underline{g}\ \underline{c}' & \underline{\Phi}(T)+\underline{\psi}(T)(\underline{b}\ \underline{k}'-\underline{g}\ \underline{c}')\end{bmatrix}\begin{bmatrix}\underline{x}(kT)\\ \hat{\underline{x}}(kT)\end{bmatrix} + \begin{bmatrix}\underline{\psi}(T)\underline{b}\\ \underline{\psi}(T)\underline{b}\end{bmatrix}u(kT) \tag{24}$$

By applying a similarity linear transformation, $\underline{P}_t$, where

$$\underline{\xi}(kT) = \underline{P}_t\begin{bmatrix}\underline{x}(kT)\\ \hat{\underline{x}}(kT)\end{bmatrix};\quad \underline{P}_t = \begin{bmatrix}\underline{I} & \underline{0}\\ \underline{I} & -\underline{I}\end{bmatrix}$$

to the system, the eigenvalues remain unchanged, and the system equations become

$$\underline{\xi}(k+1)T) = \begin{bmatrix}\underline{\Phi}(T)+\underline{h}(T)\underline{k}' & -\underline{h}(T)\underline{k}'\\ \underline{0} & \underline{\Phi}(T)-\underline{\psi}(T)\underline{g}\ \underline{c}'\end{bmatrix}\underline{\xi}(kT) + \begin{bmatrix}\underline{h}(T)\\ \underline{0}\end{bmatrix}u(kT) \tag{25}$$

$$\underline{y}(kT) = \begin{bmatrix}\underline{c}' & \underline{0}\end{bmatrix}\underline{\xi}\ (kT) \tag{26}$$

According to Gantmacher$^{(6)}$ the system matrix for (25) has a characteristic polynomial given by

$$\Delta(\lambda) = |\lambda\underline{I}-\underline{\Phi}(T)-\underline{h}(T)\underline{k}'|\,|\lambda\underline{I}-\underline{\Phi}(T)+\underline{\psi}(T)\underline{g}\ \underline{c}'|. \tag{27}$$

It has been shown in theorems (1) and (2) that the eigenvalues of $[\underline{\Phi}(T)+\underline{h}(T)\underline{k}']$ and $[\underline{\Phi}(T)-\psi(T)\underline{g}\ \underline{c}']$ are completely arbitrary. Since the characteristic equation of the compensated system, $\Delta(\lambda)$ has these eigenvalues, then the eigenvalues of both the compensated system and the asymptotic estimator compensator are arbitrary. Thus, the following theorem has been proven.

Theorem 5: Given a LSDS

$$\underline{x}((k+1)T) = \underline{\Phi}(T)\underline{x}(kT) + \underline{\psi}(T)\underline{b}u(kT) \tag{28}$$

$$y(kT) = \underline{c}'\underline{x}(kT) \tag{29}$$

which is both completely observable and completely controllable. This system can always be stabilized by using feedback from a sampled-data asymptotic estimator described by the difference equation

$$\begin{aligned}\hat{\underline{x}}((k+1)T) = {} & (\underline{\Phi}(T) - \underline{\psi}(T)\underline{g}\ \underline{c}')\hat{\underline{x}}(kT) + \underline{\psi}(T)\underline{g}\ \underline{c}'\underline{x}(kT) \\ & + \underline{\psi}(T)\underline{b}u(kT)\end{aligned} \tag{30}$$

where $\underline{g}$ can be arbitrarily selected. Applying feedback, $u(kT) = \underline{k}'\hat{\underline{x}}(kT)$ allows the arbitrary selection of system eigenvalues and the choice of $\underline{g}$ allows the arbitrary seletion of estimator eigenvalues.

Stabilization of Lur'e Type Sampled-Data Systems

Stabilization Using y_1 and y_2

Consider the system described by the set of difference equations (1) and (2). In addition assume that $[\underline{\Phi}(T),\underline{d}']$ is a completely observable pair and $[\underline{\Phi}(T),\underline{h}(T)]$ is a completely controllable pair. Now consider that an observer is constructed and placed in cascade with this system. The equations describing this system as shown in Figure (2) are

$$\underline{x}((k+1)T) = \underline{\Phi}(T)\underline{x}(kT) + \underline{\psi}(T)\underline{b}\ \underline{k}'\hat{\underline{x}}(kT) + \underline{\psi}(T)\underline{b}\theta(\sigma) \tag{31}$$

$$\begin{aligned}\hat{\underline{x}}((k+1)T) = {} & \underline{\Phi}(T)\hat{\underline{x}}(kT) + \underline{\psi}(T)\underline{b}\ \underline{k}'\hat{\underline{x}}(kT) + \underline{\psi}(T)\underline{b}\theta(\sigma) \\ & + \underline{\psi}(T)\underline{g}\ \underline{d}'\underline{x}(kT) - \underline{\psi}(T)\underline{g}\ \underline{d}'\hat{\underline{x}}(kT)\end{aligned} \tag{32}$$

$$\sigma = \underline{c}'\underline{x}(kT) \tag{33}$$

When the same transformation, $\underline{P}_t$, is used equations (32) and (33) become

$$\underline{\xi}((k+1)T) = \begin{bmatrix}\underline{\Phi}(T)+\underline{\psi}(T)\underline{b}\ \underline{k}' & -\underline{\psi}(T)\underline{b}\ \underline{k}' \\ \underline{0} & \underline{\Phi}(T)-\underline{\psi}(T)\underline{g}\ \underline{d}'\end{bmatrix}\underline{\xi}(kT) + \begin{bmatrix}\underline{\psi}(T)\underline{b} \\ \underline{0}\end{bmatrix}\theta(\sigma) \tag{34}$$

The output, σ, is given by

$$\sigma = [\underline{c}',\underline{0}]\underline{\xi}(kT) \tag{35}$$

From equations (34) and theorems (1) and (2) it is clear that the eigenvalues of the LSDS are arbitrary. Moreover, since σ is a function of the state variable, $\underline{x}(kT)$ only, then the system in (34) can be considered as two systems. One which is completly linear with eigenvalues arbitrarily determinable by selection of $\underline{g}$, and, the compensated system with a transition matrix $(\underline{\Phi}(T)+\underline{\psi}(T)\underline{b}\ \underline{k}')$ which has eigenvalues arbitrarily determinable by selection of $\underline{k}$ which also has this extra input

$\underline{\psi}(T)\underline{b}\ \underline{k}'(\underline{x}(kT) - \hat{\underline{x}}(kT)$.

The difference equations describing this new type system are

$$\underline{x}((k+1)T) = (\underline{\Phi}(T)+\underline{\psi}(T)\underline{b}\ \underline{k}')\underline{x}(kT) + \underline{\psi}(T)\underline{b}\theta(\sigma) - \underline{\psi}(T)\underline{b}\ \underline{k}'(\underline{x}(kT) - \hat{\underline{x}}(kT)). \qquad (36)$$

Which from equation (18) can be written as

$$\underline{x}((k+1)T) = (\underline{\Phi}(T) + \underline{\psi}(T)\underline{b}\ \underline{k}')\underline{x}(kT) + \underline{\psi}(T)\underline{b}\theta(\sigma) -\underline{\psi}(T)\underline{b}\ \underline{k}'(\underline{\Phi}(T) - \underline{\psi}(T)\underline{g}\ \underline{d}')^k(\underline{x}(0) - \hat{\underline{x}}(0)) \qquad (37)$$

Obviously, the last term → 0 as k → ∞ and therefore can be neglected.

Now, to study the stability of the NLSDS (nonlinear sampled-data system) by means of a sampled-data criterion analogous to the Popov criterion, the transfer function,

$G(z) = \dfrac{-\sigma(z)}{\theta_\sigma(z)}$ must be obtained.

The transfer function of the compensated system is easily shown to be

$$G_c(z) = \frac{-\sigma(z)}{\theta_\sigma(z)} = -\underline{c}'(\underline{I}z - \underline{\Phi} - \underline{\psi}\ \underline{b}\ \underline{k}')^{-1}\underline{\psi}\ \underline{b} \qquad (38)$$

Without loss of generality we can still consider that the linear part of the system, that is, the sampled data part, is in control canonical form. Thus, equation (38) is directly shown to be

$$G_c(z) = \frac{-c_n z^{n-1} - c_{n-1} z^{n-2} - \ldots\ldots - c_1}{z^n + (\nu_1 + k_n)z^{n-1} + \ldots\ldots + (\nu_n + k_1)} \qquad (39)$$

The closest analog of Popov's result for the sampled-data case is the Tsypkin criterion of Theorem 3.

In this criterion the asymptotic stability in the large was insured in the sector [0,k] if $\theta'(\sigma) \geq 0$ and

$$\mathrm{Re}[1 + \alpha(1-z^{-1})]G(z) + \frac{1}{k} > 0 \qquad (40)$$

$$|z| = 1, \ \alpha \geq 0$$

If the case where $\alpha = 0$ is considered, the criterion is reduced to

$$\mathrm{Re}\ G(z) > -\frac{1}{k} \text{ for all } |z| = 1. \qquad (41)$$

To show that equation (41) is always true for our system as long as the poles of $G_p(z)$ are arbitrary, the transfer function $G_c(z)$ is taken to the w - plane by the bilinear transformation,

$$w = \frac{z-1}{z+1} = x + jy \qquad (42)$$

Solving for z shows the relationship

$$z = \frac{1+w}{1-w} = u + jv \qquad (43)$$

By this transformation the entire unit circle in the z - plane is mapped onto the entire left half of the w - plane, and the frequency response can be obtained by setting w = jv. Thus, equation (39) can be written in the following form,

$$G_c(w) = -c_n \frac{\prod_{i=1}^{n-1} (1+\beta_i)(1-w) \prod_{i=1}^{n-1} \left(1+ \frac{w}{(1+\beta_i)/(1-\beta_i)}\right)}{\prod_{j=1}^{n} (1+\alpha_j) \prod_{i=1}^{n} \left(1+ \frac{w}{(1+\alpha_j)/(1-\alpha_j)}\right)} \tag{44}$$

where

$$G_c(w) = G_c\left(z = \frac{1+w}{1-w}\right) \tag{45}$$

Let $A_j = \frac{1+\alpha_j}{1-\alpha_j}$, and $B_i = \frac{1+\beta_i}{1-\beta_i}$. Substituting in (44) yields (46)

$$G_c(w) = \frac{-c_n \prod_{i=1}^{n-1} (1+\beta_i)(1-w) \prod_{i=1}^{n-1} \left(1 + \frac{w}{B_i}\right)}{\prod_{j=1}^{n} (1+\alpha_j) \prod_{j=1}^{n} \left(1 + \frac{w}{A_j}\right)} \tag{47}$$

Setting $w = jv$, equation (41) becomes

$$\operatorname{Re} G(iv) > -1/k \tag{48}$$

$$G_c(jv) = \frac{-c_n \prod_{i=1}^{n-1} (1+\beta_i)(1-jv) \prod_{i=1}^{n-1} \left(1+ \frac{jv}{B_i}\right)}{\prod_{j=1}^{n} (1+\alpha_j) \prod_{j=1}^{n} \left(1+ \frac{jv}{A_j}\right)} \tag{49}$$

Where β_i are the fixed zeros in the w plane and the A_i are the arbitrary poles in the w plane. Equation (41) can always be achieved by placing the arbitrary poles such that

$$|G_c(jv)| < \frac{1}{k} \tag{50}$$

To show this let us write (47) in the following factored form

$$G_c(jv) = \frac{-c_n(1+\beta_1)(1+\beta_2)\ldots\ldots(1+\beta_k)\left(1+ \frac{jv}{B_1}\right)\left(1+ \frac{jv}{B_2}\right)\ldots\ldots\ldots\ldots\ldots}{(1+\alpha)_1(1+\alpha_2)\ldots\ldots(1+\alpha_n)\left(1+ \frac{jv}{A_1}\right)\left(1+ \frac{jv}{A_2}\right)\ldots\left(1+ \frac{jv}{A_k}\right)\ldots\ldots}$$

$$\frac{\ldots\ldots\left(1+ \frac{jv}{B_k}\right)}{\ldots\ldots\left(1+ \frac{jv}{A_n}\right)} \tag{51}$$

Where, again, B_i are the fixed zeros and the A_i are the arbitrary poles of $G_c(jv)$. If the values of A_i, $i = 1,2,3\ldots,k$, are chosen such that

$$A_i = -|\operatorname{Re} B_i| + j \operatorname{Im} B_i - \varepsilon \tag{52}$$

where ε is an arbitrary small positive constant being added to maintain the complete controllability of $[\underline{\phi}(T)+\underline{\psi}(T)\underline{b}\ \underline{k}', \underline{\psi}(T)\underline{b}]$, then $G_c(jv)$ can be written as

$$G_c(jv) = \frac{-c_n(1+\eta(\varepsilon))\underline{|\,\theta(v)}}{(1+\alpha_{k+1})(1+\alpha_{k+2})\ldots\ldots(1+\alpha_n)\left(1+ \frac{jv}{A_{k+1}}\right)\left(1+ \frac{jv}{A_{k+2}}\right)\ldots\left(1+ \frac{jv}{A_n}\right)} \tag{53}$$

where $\eta(\varepsilon)$ is a finite constant which is a result of the incomplete cancellation of the zeros because of the addition of the constant, ε. The angle $\delta(v)$ appeared here as a result of the constant, ε, and the possible cancellation of the zeros with a positive real part with a negative real part. Now, if all the rest of the arbitrary poles A_i $(i = k+1, k+2, \ldots, n)$ are selected to be real to prevent resonance effects, the magnitude of $G_c(jv)$ is bounded by

$$|G_c(jv)| \leq \left| \frac{-c_n(1+\eta(\varepsilon))\underline{|\delta(v)}}{(1+\alpha_{k+1})(1+\alpha_{k+2})\ldots\ldots\ldots\ldots(1+\alpha_n)} \right| \tag{54}$$

Obviously, there is always at least one arbitrary $(1+\alpha_n)$, which can make $G_c(jv)$ less than $\frac{1}{k}$. Thus, the following theorem can be stated.

Theorem 6 : Consider a NLSDS, described by the difference equations

$$\underline{x}((k+1)T) = \underline{\Phi}(T)\underline{x}(kT) + \underline{h}(T)\theta(\sigma) \tag{55}$$

$$\sigma = \underline{c}'\underline{x}(kT) \tag{56}$$

where $\theta(\sigma)$ satisfies the conditions

$$\theta(\sigma) = 0 \qquad \text{for } \sigma = 0 \tag{57}$$

$$0 \leq \frac{\theta(\sigma)}{\sigma} \leq k, \ \theta'(\sigma) \geq 0 \quad \text{for } \sigma \neq 0 \tag{58}$$

$$\lim_{\sigma\to\infty} \int_0^\sigma \theta(\sigma)d\sigma \to \infty \tag{59}$$

Assume that the output, $y_2(kT) = \underline{d}'\underline{x}(kT)$, is completely observable. Moreover, assume that $\underline{h}(T)$ provides complete controllability of the linear portion of the system. Then, a sampled-data asymptotic estimator can be constructed whose inputs are $y_1(kT) = \theta(\sigma)$ and $y_2(kT) = \underline{d}'\underline{x}(kT)$ and whose output state is an asymptotic estimate of state $\underline{x}(kT)$, (i.e., $(\underline{x}(kT) - \hat{\underline{x}}(kT)) \to 0$ as $k \to \infty$) and where the dynamics (i.e., the eigenvalues of $\underline{M}$ in $\underline{x}(kT) - \hat{\underline{x}}(kT) = \underline{M}^k(\tilde{\underline{x}}(0))$ are arbitrary. In addition the output of the estimator can always be used to provide feedback which will make the system absolutely stable for all $\theta(\sigma)$ described above.

SUMMARY

It has been shown that a sampled-data estimator can always be constructed which gives an estimated vector, $\hat{\underline{x}}(kT)$ plus an error; the dynamics of the error term are completely arbitrary. This estimator can always be utilized to stabilize a linear sampled data system provided that the system is completely controllable and completely observable. In addition it can be used to stabilize a Popov type nonlinear sampled-data system in the sector $[0,k]$ where k is finite, if both the nonlinear function and a completely observable output of the linear portion are available as inputs to the estimator.

ACKNOWLEDGEMENTS

M. I. Younis wishes to acknowledge with thanks support from the University of Pittsburgh Space Research Coordination Center under NASA Grant NSG 416. W. G. Vogt wishes to acknowledge with thanks research support from the Gulf Oil Corporation.

REFERENCES

1. D. G. Luenberger, "Observing the State of a Linear System," IEEE Transactions on Military Electronics, V. Mil - 8, pp. 74-80, April, 1964.

2. R. E. Strane and W. G. Vogt, "Stabilization of a Class of Nonlinear Systems Using a Linear System Estimator, Proc. Seventh Annual Allerton Conference on Circuit and System Theory, pp. 313-321, October, 1969.

3. M. I. Younis, "Stabilization of Nonlinear Sampled-Data System Using a Linear Sampled-Data System Estimator," Ph. D. Dissertation in progress, Dept. of Electrical Engineering, University of Pittsburgh, Pittsburgh, Pa.

4. R. E. Kalman, P. E. Falb and M. I. Arbib, Topics in Mathematical System Theory, (New York: McGraw-Hill Book Company), 1969.

5. Y. Z. Tsypkin, "A Criterion for Absolute Stability of Automatic Pulse Systems with Monotonic Characteristics of the Nonlinear Element", Dokl. Akad. Nauk. (USSR), Vol. 155, pp. 1029-1032.

6. F. R. Gantmacher, The Theory of Matrices, Vol. I. New York: Chelsea Publishing Co., 1959.

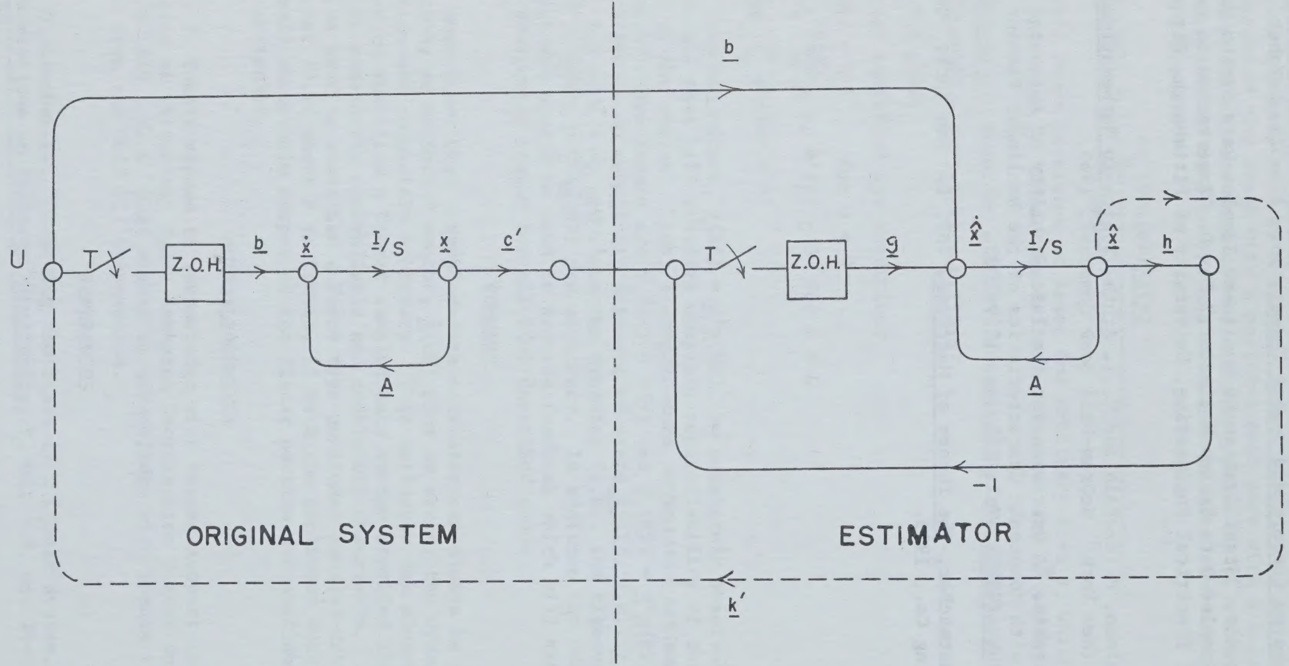

U
T
Z.O.H.
$\underline{b}$
$\dot{\underline{x}}$
$\underline{I}/S$
$\underline{x}$
$\underline{c}'$
$\underline{A}$
ORIGINAL SYSTEM
$\underline{b}$
T
Z.O.H.
$\underline{g}$
$\dot{\hat{\underline{x}}}$
$\underline{I}/S$
$\hat{\underline{x}}$
$\underline{h}$
$\underline{A}$
-1
ESTIMATOR
$\underline{k}'$

Figure 1. Signal Flow Diagram of the LSDS Cascaded by a Sampled-Data Estimator

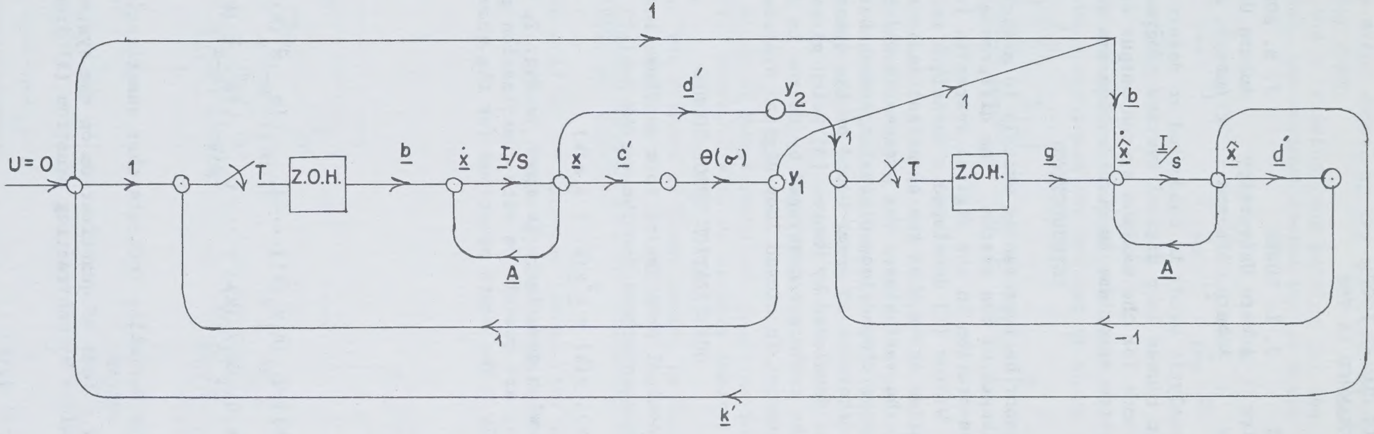

U=0
1
T
Z.O.H.
b
ẋ
I/s
x
c′
θ(σ)
d′
y2
y1
1
1
T
Z.O.H.
g
x̂̇
I/s
x̂
d′
b
1
A
1
A
-1
k′

Figure 2. Signal Flow Diagram - Lur'e Type Sampled-Data System and Estimator for Stabilization

QUANTIZATION ERROR IN CLOSED LOOP DIGITAL SYSTEMS PROGRAMMED IN THE CANONICAL FORM*

MICHAEL H. FONG
Auburn University
Auburn, Alabama

J. L. LOWRY
Auburn University
Auburn, Alabama

J. S. BOLAND, III
Auburn University
Auburn, Alabama

Abstract - An analytic study is conducted to determine the quantization error present in closed loop digital systems programmed in the canonical form. Upper bounds for the maximum system output error and the maximum steady state system error due to quantization are derived.

INTRODUCTION

Most numbers cannot be represented exactly in a digital system because of the finite word length of the system. The difference between the exact number and its representation in the digital system is referred to as the quantization error. Widrow [1] developed a technique to estimate the root-mean-square quantization error when the quantization is small relative to the dynamic range of the variables. For systems in which this condition is not met, such as in some closed loop digital systems, Bertram [2] uses the concept of state to determine an upper bound on the quantization error. A simpler technique was developed by Monroe [3] which gives equivalent results. This paper extends the technique developed by Monroe to determine the quantization error present in closed loop digital systems programmed in the canonical form.

QUANTIZATION ERROR BOUNDS

Consider the canonical programming form as shown in Fig. 1 for m=n. The corresponding state equations for the system are

$$\underline{x}(k+1) = A\underline{x}(k) + \underline{b}e(k), \; y(k) = \underline{c}^T\underline{x}(k) + a_n e(k) \tag{1}$$

The same system with quantizers is shown in Fig. 2, where the blocks labeled Q_{10}, Q_{11}, . . ., are quantizers with quantization granularity h_{10}, h_{11}, . . ., respectively. The state equations for the quantized system become

$$\bar{x}_1(k+1) = \bar{x}_2(k)$$

$$\bar{x}_2(k+1) = \bar{x}_3(k)$$

$$\vdots$$

$$\bar{x}_{n-1}(k+1) = \bar{x}_n(k)$$

$$\bar{x}_n(k+1) = -Q_{10}[b_o\bar{x}_1(k)]-Q_{11}[b_1\bar{x}_2(k)]\cdots-Q_{1(n-1)}[b_{n-1}\bar{x}_n(k)] + Q_{1n}[e(k)].$$

$$\bar{y}(k) = Q_{20}[a_o\bar{x}_1(k)] + Q_{21}[a_1\bar{x}_2(k)] + \cdots Q_{2(n-1)}[a_{n-1}\bar{x}_n(k)]$$
$$+ Q_{2n}\{a_n[\bar{x}_n(k+1)]\}. \tag{2}$$

The bar over the state variables indicate that quantization effects have been included in the system.

To determine the effect of quantization on the system, a set of error equations can be obtained by subtracting equation (1) from equation (2) and noting that

$$Q_i[b_j\bar{x}_j(k)] = b_j\bar{x}_j(k) + q_i(k) \tag{3}$$

where $q_i(k)$ is the error caused by the quantizer Q_i operating on $b_j\bar{x}_j(k)$ at the sampling instant $t=kT$.

After taking the z transform and solving for the corresponding maximum time domain function, the maximum system output error for truncation is given by

$$y_\gamma(k)_{max} = \sum_{j=o}^{n} h_{1j} \sum_{i=o}^{k} g_j(i) + a_n \sum_{i=o}^{n} h_{1i} + \sum_{i=o}^{n} h_{2i} \tag{4}$$

where $\underline{G}^T = \underline{C}^T [ZI-A]^{-1}S.$ (5)

The steady state output error can be obtained by applying the final value theorem to the error equations and can be shown to be

$$y_{\gamma_s} = \frac{\left(\sum_{n=o}^{n-1} c_i\right)\left(\sum_{i=o}^{n} h_{1i}\right)}{1 + \sum_{i=o}^{n-1} b_i} + a_n \sum_{i=o}^{n} h_{1i} + \sum_{i=o}^{n} h_{2i} \tag{6}$$

It can be seen that the resulting bounds are independent of system excitation. The bounds developed may be used to estimate the necessary encoding equipment and the maximum computer word length required to achieve a given system specification.

REFERENCES

[1] B. Widrow, "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," Trans. I.R.E. on Circuit Theory, pp. 266-276, December 1956.

[2] J. E. Bertram, "The Effects of Quantization in Feedback Systems," AD 154 227, March 1958.

[3] A. J. Monroe, Digital Processes for Sampled-Data Systems, New York: McGraw-Hill, 1959, pp. 362-381.

*This work was supported by the National Aeronautics and Space Administration, Huntsville, Alabama, under contracts NAS8-20004 and NAS8-20104.

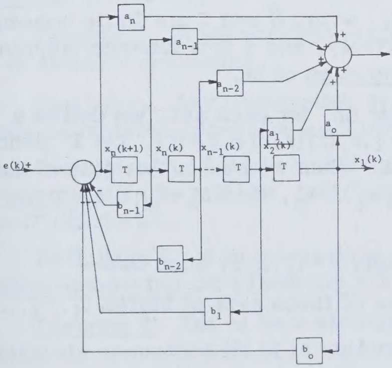

a_n
a_{n-1}
a_{n-2}
a_1
a_o
y(k)
e(k)+
$x_n(k+1)$
$x_n(k)$
$x_{n-1}(k)$
$x_2(k)$
$x_1(k)$
T
b_{n-1}
b_{n-2}
b_1
b_o

Fig. 1 Canonical Programming Form without Quantizers.

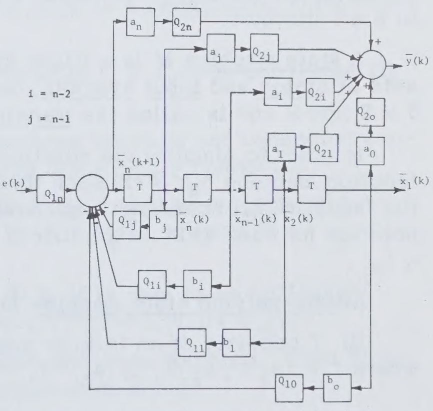

a_n
Q_{2n}
a_j
Q_{2j}
$\bar{y}(k)$
i = n-2
j = n-1
a_i
Q_{2i}
Q_{2o}
a_1
Q_{21}
a_0
$\bar{x}_n(k+1)$
e(k)
Q_{1n}
T
$\bar{x}_1(k)$
Q_{1j}
b_j
$\bar{x}_n(k)$
$\bar{x}_{n-1}(k)$
$\bar{x}_2(k)$
Q_{1i}
b_i
Q_{11}
b_1
Q_{10}
b_o

Fig. 2 Canonical Programming Form with Quantizers.

ON PERIODICITY OF SEQUENTIAL MACHINES

RATAN K. GUHA and RAYMOND T. YEH
Department of Computer Sciences and Electronics Research Center
The University of Texas at Austin, Austin, Texas

ABSTRACT

It was shown by Gill and Flexer [3] that the necessary and sufficient condition for the existence of non-trivial periodic decomposition of a sequential machine corresponds to the existence of a non-trivial cyclic partition. We have, in this paper, characterized the existence or nonexistence of cyclic partitions of machines under various connectedness conditions. Also, their theory is generalized here utilizing the concept of cyclic covers. As a consequence of this generalization, the open problem posed by Gill and Flexer [3] is solved. Upper bounds of periodicity of sequential machines are obtained using Sperner's theorem [4]. The bound is exact for transient-free machines.

I. PRELIMINARIES

A family of subsets $C = \{S_i\}_{i \in I}$, I an index set, on a set S is called a cover of S iff $\bigcup_{i \in I} S_i = S$. We say this cover is irredundant if $(\forall i, j \in I)[S_i \subseteq S_j \rightarrow S_i = S_j]$. We define Max (C) as the set $\{S_i \in C \mid (\forall S_j \in C)[S_i \neq S_j \rightarrow S_i \not\subseteq S_j]$. We note that if C is irredundant, then Max (C) = C. If $C_1 = \{S_i\}_{i \in I}$ and $C_2 = \{T_j\}_{j \in J}$ are two covers on a set S, then C_1 is said to be a refinement of C_2 (in symbols $C_j \leq C_2$) iff $(\forall S_i \in C_1)(\exists T_j \in C_2)[S_i \subseteq T_j]$. It is easily seen that $C_1 = C_2$ iff $C_1 \leq C_2$ and $C_2 \leq C_1$.

A relation θ on a set S is said to be a compatibility relation iff θ is both reflexive and symmetric. We note that each compatibility relation $\theta \subseteq S^2$ defines a maximal irredundant cover C_θ on S such that

$$C_\theta = \{T \subseteq S \mid (\forall s \in S)[s \in T \rightleftarrows (\forall t \in T)[(s, t) \in \theta]]\}$$

A graph G is a pair (V, ρ), where V is a set of vertices and ρ is a relation on V. A path Π on G is a finite sequence $V_0, V_1, \ldots, V_n$ of elements of such that $(\forall 0 \leq i < n)[(V_i, V_{i+1}) \in \rho]$. Π is called a circuit if $V_0 = V_n$, and n is called the period of the circuit. Π is called a simple circuit if furthermore all elements in Π are distinct.

A state machine M is a triple (S, Σ, δ), where S and Σ are finite nonempty sets of states and input symbols, respectively, and δ is a function mapping $S \times \Sigma$ into S and is called the transition function of M.

In order to simplify the notations later on, for each $a \in \Sigma$, we define a function mapping $\delta_a : S \rightarrow S$ such that $\delta_a = \{(s, t) \mid (s, a) \circ \delta = t)\}$. Let Σ^+ denote the family of all nonempty sequences of Σ. Then we can further extend the notation for each $x \in \Sigma^+$, such that if $x = a_1 a_2 \ldots a_n$, then $\delta_x = \delta_{a_1} \circ \delta_{a_2} \circ \ldots \circ \delta_{a_n}$.

A time-varying state machine is a triple $A = (\Gamma, \Sigma, \Delta)$ such that

(i) Γ consists of an infinite sequence of finite sets of states $\Gamma_1, \Gamma_2, \ldots$, where $\Gamma_i = \{s_1^{(i)}, s_2^{(i)}, \ldots, s_{n_i}^{(i)}\}$.

(ii) Σ is a finite set of input symbols.

(iii) Δ consists of an infinite sequence of transition functions $f^{(1)}, f^{(2)}, \ldots,$ such that $f^{(i)} : \Gamma_i \times \Sigma \rightarrow \Gamma_{i+1}$.

Following the notations above we define functions $f_x^{(k)}$, $k > 0$, $x \varepsilon \Sigma^+$, by the following rules:

(i) $f_a^{(k)} = \{(s_i^{(k)}, s_j^{(k+1)}) \mid (s_i^{(k)}, a) \circ f^{(k)} = s_j^{(k+1)}\}$.

(ii) If $x = a_1 a_2 \ldots a_n \varepsilon \Sigma^+$, then

$$f_x^{(k)} = f_{a_1}^{(k)} \circ f_{a_2}^{(k+1)} \circ \ldots \circ f_{a_n}^{(k+1)} .$$

If $A = (\Gamma, \Sigma, \Delta)$ and $A' = (\Gamma', \Sigma, \Delta')$ are two time-varying state machines over the same input alphabet Σ, then we say A' is a sub-machine of A iff

$$(\forall i > 0)[\Gamma_i' \subseteq \Gamma_i \wedge f^{(i)'} = f^{(i)} \mid \Gamma_i'].$$

A time varying state machine $A = (\Gamma, \Sigma, \Delta)$ is said to be periodic iff

$$(\exists \zeta \geq 0)(\exists T > 0)[(\forall k > o)[S_{k+T+\zeta} = S_{k+\zeta} \wedge f^{(k+T+\zeta)} = f^{(k+\zeta)}]].$$

We call T the period of A. If $\zeta = 0$, then A is said to be strictly periodic. Since we can always combine the state sets of a periodic machine to yield a strictly periodic machine, we will only consider strictly periodic machines in this paper. However, we will simply write periodic machine instead of strictly periodic machine.

II. PERIODIC DECOMPOSITION OF STATE MACHINES USING CYCLIC PARTITIONS

Definition 1. A state machine $M = (S, \Sigma, \delta)$ is said to be connected iff

$$(\forall s, t \varepsilon S)(\exists n > 0)(\exists s_1 = s, s_2, \ldots, s_n = t)[(\forall 1 \leq i < n)$$
$$[(s_i, s_{i+1}) \varepsilon \delta_a \vee (s_{i+1}, s_i) \varepsilon \delta_a]].$$

Definition 2. A state machine $M = (S, \Sigma, \delta)$ is said to be strongly connected iff $(\forall s, tS)(\exists s \varepsilon \Sigma^+)[(s, t) \varepsilon \delta_x]$.

Definition 3. A state machine $M = (S, \Sigma, \delta)$ is a permutation state machine iff $(\forall a \varepsilon A)[|S \circ \delta_a| = |S|]$.

Theorem 1. A permutation state machine $M = (S, \Sigma, \delta)$ is connected iff it is strongly connected.

Corollary. Any permutation state machine which is not connected consists of isolated sub-state machines which are strongly connected.

Definition 4. A state machine $M = (S, \Sigma, \delta)$ is said to be uniformly connected iff $(\exists p > 0)[(\forall s, t \varepsilon S)(\exists x \varepsilon \Sigma^{(p)})[(s, t) \varepsilon \delta_x]]$, where $\Sigma^{(p)} = \{x \varepsilon \Sigma^+ \mid lg(x) = p\}$.

Definition 5. A state machine $M = (S, \Sigma, \delta)$ is said to be uniformly initial-connected iff $(\exists s \varepsilon S)(\exists p > 0)\ [(\forall t \varepsilon S)(\exists x \varepsilon \Sigma^{(p)})[(s, t) \varepsilon \delta_x]]$.

Theorem 2. Let M be a strongly connected state machine. Then M is uniformly connected iff M is uniformly initial-connected.

Theorem 3. A state machine $M = (S, \Sigma, \delta)$ is uniformly connected iff M is strongly connected and

$$(\exists s^* \epsilon S)\,(\exists k > 0)\,(\exists x \epsilon \Sigma^{(k)})\,[(s^*, s^*) \epsilon \delta_x] \wedge (\forall s \epsilon S)\,(\exists m > 0)\,(\exists x \epsilon \Sigma^{(mk)})\,[(s^*, s) \epsilon \delta_x].$$

Corollary. If $M = (S, \Sigma, \delta)$ is a strongly connected state machine such that $(\exists s \epsilon S)\,(\exists a \epsilon \Sigma)\,[(s, s) \epsilon \delta_a]$, then M is uniformly connected.

Definition 7. Let $M = (s, \Sigma, \delta)$ be a state machine and $A = (\Gamma, \Sigma, \Delta)$ a periodic machine. We say A realizes M iff there is a sub-machine $A' = (\Gamma', \Sigma\ \Delta')$ of A and a mapping $\varphi: \Gamma' \rightarrow S$ such that φ is one-to-one and onto, and $(\forall i > 0)\,(\underline{\forall} a \epsilon \Sigma)\,[f_a^{(i)'} \circ \varphi = \varphi \circ \delta_a]$. In this case, we say A is a periodic decomposition of M. We say it is a non-trivial periodic decomposition if $(\forall S_i \epsilon \Gamma)\,[|S_i| < |S|]$.

Theorem 4. A state machine $M = (S, \Sigma, \delta)$ has a nontrivial periodic decomposition iff there exists a cyclic partition of order greater than one on M.

Theorem 5. A uniformly connected state machine has no non-trivial periodic decomposition.

Proof: Let us suppose that M has a nontrivial periodic decomposition, i.e., there exists a partition $S_0, S_1, \ldots, S_{k-1}$ on S such that:

$$(\forall m > 0)\,(\forall 0 \le i < k)\,[\bigcup_{x \epsilon \Sigma^{(m)}} S_i \circ \delta_x \subseteq S_{i+m\,(\mathrm{mod}\,k)} \wedge S_i \subsetneq S]. \tag{1}$$

Since M is uniformly connected, so $(\forall m > p)\,(\forall s \epsilon S)\,[\bigcup_{x \epsilon \Sigma^{(m)}} s \circ \delta_x = S]$. This is a contradiction. Hence M cannot have nontrivial periodic decomposition.

Theorem 6. Any strongly connected but not uniformly connected permutation state machine $M = (S, \Sigma, \delta)$ can be realized by a periodic state machine with period $T\,(T > 1)$ such that T is a divisor of $|S|$. Furthermore, the number of states of the periodic machine at any time will be $|S|/T$.

Proof: Let s^* be an arbitrary element of M. We construct the sequence of sets $\{S_0, S_1, S_2, \ldots\}$ as follows:

(i) $S_0 = \{s^*\}$;

(ii) $(\forall i > 0)\,[S_i = \bigcup_{x \epsilon \Sigma^{(i)}} s^* \circ \delta_x]$

We see that every set S_i is a proper subset of S. If every S_i contains a single element, then the theorem is obviously true. Thus, we may now assume that there exists an S_i such that $|S_i| > 1$. Since M is a permutation machine, we must have

$$|S_0| \le |S_1| \le \ldots \le |S_n| \le \ldots \tag{2}$$

Since S is a finite set and every S_i is a proper subset of S, it follows from (2) and the fact that $|S_i| > 1$ for some S_i that

$$(\exists m \ge 1)\,[(\forall j \ge 1)\,[|S_m| = |S_{m+j}|] \wedge |S_{m-1}| < |S_m|].$$

Thus, (2) becomes

$$|S_0| \leq |S_1| \leq \ldots \leq |S_{m-1}| \leq |S_m| = |S_{m+1}| = \cdots = |S_{m+j}| = \cdots$$

It is also easy to see that the following hold:

(iii) $(\forall k \geq m)\,[S_k \cap S_{k+1} = \emptyset]$.

(iv) $(\forall k \geq m)\,(\forall i \geq 2)\,[S_k \cap S_{k+1} = \emptyset \vee S_k = S_{k+i}]$.

Let us now consider the subsets $S_m, S_{m+1}, \cdots, S_{m+T-1}$, where T is the smallest integer such that $S_{m+T} = S_m$. Since S is finite and M strongly connected, such a T exists. It is clear from (iii) that $T > 1$. Furthermore, M is strongly connected and implies that $\bigcup_{i=0}^{T-1} S_{m+1} = S$. We now see that

$$(\forall 0 \leq i, j \leq T-1)\,[i \neq j \rightarrow S_{m+i} \neq S_{m+j}] \tag{3}$$

(3) together with (iii) and (iv) implies that $(\forall 0 \leq i, j \leq T-1)\,[i \neq j \rightarrow S_{m+i} \cap S_{m+j} = \emptyset]$. Thus, the subsets $\{S_m, S_{m-1}, \cdots, S_{m+T-1}\}$ define a cyclic Partition of order $(T > 1)$ on M. Furthermore, each class in the partition has the same number of states. Therefore, the theorem follows from theorem 4.

Corollary. If M is a strongly connected but not uniformly connected permutation state machine which has a prime number of states, then M has a periodic representation whose state set at each instant is a singleton set.

We observe here that if a machine is connected but not strongly connected, then it may or may not have a non-trivial periodic decomposition.

III. GENERALIZED DECOMPOSITION AND THE BASIC COVER OF A STATE MACHINE

We observe at this point that it is possible to enlarge a periodic indecomposable machine to one which has non-trivial periodic decomposition using state-splitting technique [1]. Thus, the concept of periodic decomposition using cyclic partitions seems to be too narrow, and it seems that the concept of cyclic cover is a natural generalization. In the sequel, we shall refer to the concepts of realization and periodic decomposition defined in the last section as simple realization and simple periodic decomposition, respectively. We shall now proceed to generalize these notions based on the concept of cyclic cover.

Definition 8. Let $M = (S, \Sigma, \delta)$ be a state machine. A cover $C = \{S_0, S_1, \cdots, S_{k-1}\}$ on S is called a cyclic cover of order k on M iff

$$(\forall 0 \leq i < k)\,\left[\bigcup_{a \in \Sigma} S_i \circ \delta_a \subseteq S_{i+1\,(\mathrm{mod}\,k)}\right]$$

Definition 9. Let $M = (S, \Sigma, \delta)$ be a state machine and $A = (\Gamma, \Sigma, \Delta)$ a periodic machine. We say A realizes M iff there is a sub-machine $A' = (\Gamma', \Sigma, \Delta')$ of A and a mapping $\varphi: \Gamma' \rightarrow S$ such that φ is onto and

$$(\forall i > 0)\,(\forall a \in \Sigma)\,[f_a^{(i)} \circ \varphi = \varphi \circ \delta_a]$$

In this case, we say A is a periodic decomposition of M. It is called a non-trivial periodic decomposition of M if furthermore

(i) $(\forall \Gamma_i \in \Gamma)\,[\,|\Gamma_i| < |S|\,];$

(ii) $(\forall \Gamma'_i, \Gamma'_j \in \Gamma')\,[i \neq j \rightarrow \Gamma'_i \circ \varphi \not\subseteq \Gamma'_j \circ \varphi \wedge \Gamma'_j \circ \varphi \not\subseteq \Gamma'_i \circ \varphi]$

Theorem 7. A state machine $M = (S, \Sigma, \delta)$ has a nontrivial periodic decomposition iff there exists an irredundant cyclic cover of order greater than one on M.

Definition 10. Let $M = (S, \Sigma, \delta)$ be a state machine. We define a compatibility relation $\theta(M)$ on S as follows:

(i) $\theta_0 = \{(s,s) \mid s \in S\}$;

(ii) $\theta_{i+1} = \theta_i \cup \{(s,t) \mid (\exists a, b \in \Sigma)\,[(s,t) \in \delta_a^{-1} \circ \theta_i \circ \delta_b]\},\ i \geq 0;$

(iii) $\theta(M) = \theta_p$, where p is the smallest integer such that $\theta_p = \theta_{p+1}$.

It is easily seen from the above definition that $\theta(M)$ is a well defined compatibility relation since S is finite. Furthermore, we have

Lemma 1. Let $M = (S, \Sigma, \delta)$ be a state machine. Then the corresponding cover $C_{\theta(M)}$ of $\theta(M)$ satisfies the following condition:

$$(\forall \alpha \in C_{\theta(M)})\,(\exists n \geq 0)\,(\exists s \in S)\,[\alpha = \bigcup_{x \in \Sigma^{(n)}} s \circ \delta_x].$$

We will call $C_{\theta(M)}$ the basic cover of M and denote it by C_M for simplicity.

Lemma 2. If C is any irredundant cyclic cover on a state machine $M = (S, \Sigma, \delta)$, then $C_M \leq C$.

IV. PERIODICITY OF STATE MACHINES WITH DETERMINISTIC COVER GRAPHS

Definition 11. The cover graph, G_M, of a state machine M is defined to be the pair (C_M, ρ_M), where C_M is the basic cover of M and

$$\rho_M = \{(\alpha, \beta) \in C_M^2 \mid \bigcup_{\alpha \in \Sigma} \alpha \circ \delta_a \subseteq \beta\}.$$

G_M is said to be deterministic if there is exactly one outgoing edge from each vertex.

Definition 12. A state machine M is said to be transient-free if it does not contain any transient state s such that $(\forall x \in \Sigma^+)\,[(s,s) \notin \delta_x]$.

Lemma 3. If $M = (S, \Sigma, \delta)$ is a transient-free state machine, then G_M is covered by m disjoint simple circuits of G_M for some $m \geq 1$.

Theorem 8. If $M = (S, \Sigma, \delta)$ is a transient-free state machine, then the maximal period of irredundant cyclic covers, which are also covers of C_M, on M equals to the l.c.m. of periods of all simple circuits of G_M.

Definition 13. A vertex v in a graph G is said to be singly-connected to a circuit C in G if there is precisely one path from v to an element of C.

Lemma 4. Any deterministic graph G contains m disjoint simple circuits for some $m \geq 1$, and each vertex in G not belonging to any circuit is simply-connected to exactly one circuit.

It follows from lemma 4 that the vertex set of $G_M = (C_M, \rho_M)$ can be partitioned into two classes E and F, respectively, such that E contains vertices belonging to some circuit of G_M and F contains vertices not belonging to any circuit of G_M. Since all circuits are disjoint in G_M, let us assume now for simplicity that G_M is a deterministic graph containing exactly one simple circuit $L=\{\alpha_1,\cdots,\alpha_n\}$. Let $\overline{C}$ be an arbitrary irredundant cyclic cover on S (which is also a cover of C_M by assumption). In order to estimate the maximal period of any irredundant cyclic cover on S, we observe that if an element of $\overline{C}$ contains m elements of E, then every element of $\overline{C}$ must also contain m elements of E. In this case, the period of $\overline{C}$ is clearly bounded by the maximum number of distinct subsets of C_M each containing an m-element subset of E together with some elements of F. However, not every subset of F can be combined with an m-element subset of E to form a class in $\overline{C}$. In order to find all the possible elements of F which can be combined with a m-element subset of E to generate classes in $\overline{C}$, let the family of all distinct m-element subsets of E contained in elements of $\overline{C}$ be $Q=\{Q_1,\cdots,Q_\ell\}$, where $\ell \le n$. We note that Q forms a cyclic cover on E. Now for each $1 \le i \le \ell$, we define a subset P_i of F as follows:

$$P_i = \{\, \alpha\epsilon F \mid (\exists\, \beta\epsilon Q_i)\, (\exists k \ge 0)\, [\alpha \circ \rho_M^{k\ell} = \beta]\}.$$

Clearly, $\Pi=\{P_1,\cdots,P_\ell\}$ forms a cover of F (some P_i may be empty). We observe that P_i consists of precisely those elements of F which can be combined with Q_i to generate possible elements of $\overline{C}$. In particular, if $P_i=\emptyset$, then Q_i must be an element of $\overline{C}$. Hence, it follows from Sperner's theorem [4] that the maximal number of distinct elements of $\overline{C}$ which can be formed from Q_i and elements of P_i is bounded by

$$u_i = \begin{cases} \displaystyle\max_{0 \le j \le |P_i|} \binom{|P_i|}{j}, & \text{if } |P_i| \neq 0 \\[2ex] 1, & \text{otherwise.} \end{cases}$$

But this immediately implies that the period of $\overline{C}$ is bounded by $n \times \min\limits_{1 \le i \le \ell}(u_i)$. This result is summarized in the following lemma.

Lemma 5. Let M be a state machine such that G_M is a deterministic graph containing exactly one circuit. Let $\overline{C}$ be any irredundant cyclic cover on S such that each element of $\overline{C}$ contains m elements of E, where E, F, Π, Q and u_i $(1 \le i \le \ell)$ are defined above, then the period of $\overline{C}$ is bounded by $n \times \min\limits_{1 \le i \le \ell}(u_i)$.

Lemma 5 is readily extendable to the general case using lemma 4.

Theorem 9. Let $M=(S,\Sigma,\delta)$ be a state machine such that G_M is a deterministic graph containing k circuits, $L_1,\cdots,L_k$, for some $k \ge 1$. Let $\overline{C}$ be any irredundant cyclic covers on S such that each element of $\overline{C}$ contains m_i elements of L_i, $1 \le i \le k$. Then the period of $\overline{C}$ is bounded by l.c.m. $(P_1,\cdots,P_k)$ where for each $1 \le i \le k$, P_i is the bound of cyclic cover containing m_i elements of L_i given by lemma 5.

V. BOUNDS ON PERIODICITY OF STATE MACHINES WITH NON-DETERMINISTIC COVER GRAPHS

We shall adopt the convention in the sequel such that all multiple outgoing edges from a vertex in a cover graph will be denoted by dotted arrows. Furthermore, all state machines considered in this section are assumed to have non-deterministic cover graphs.

Let M be a given state machine. There are three possibilities: (1) None of the dotted arrows in G_M belong to any circuit in G_M; (2) Each dotted arrow belongs to some circuit; (3) Some of the dotted arrows belong to some circuits and others do not. These three cases are illustrated in the example below.

Example. Transition tables of machines $M_i (i=1,2,3)$ are given in tables 1,2,3 and their corresponding cover graphs are given in figs. 1,2,3 below. Clearly, each graph G_{M_i} illustrates the ith $(i=1,2,3)$ case above.

δ_1	a_1	a_2	a_3
1	4	5	5
2	7	4	5
3	6	5	4
4	9	9	9
5	10	10	10
6	11	11	11
7	9	8	8
8	4	4	4
9	7	7	7
10	5	5	5
11	12	12	12
12	6	6	6
13	1	1	1

Table 1. Transition table of M_1

δ_2	a_1	a_2
1	2	2
2	7	7
3	8	8
4	9	9
5	2	3
6	4	2
7	1	1
8	10	10
9	13	13
10	11	11
11	12	12
12	13	13
13	2	2

Table 2. Transition table of M_2

δ_3	a_1	a_2
1	2	3
2	4	4
3	5	5
4	6	6
5	7	7
6	2	2
7	2	2
8	2	9
9	10	10
10	11	11
11	9	9

Table 3. Transition table of M_3

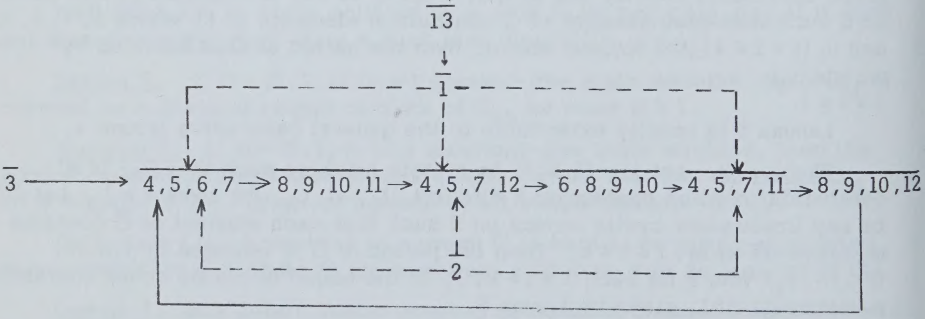


Figure 1. The cover graph of M_1

The three possibilities will now be analyzed. It will be shown that in case 1 the cover graph can be reduced to a deterministic case by means of state-splitting techniques [1], and that the cover graph in case 3 can be transformed to a set of disjoint graphs of the types 1 and 2, again by splitting states. Thus, we need only to consider case 2. We will, nevertheless, indicate the transformation for cover graphs of cases 1 and 3.

Consider the graph G_M, given in fig. 1. Since both vertices $\bar{1}$ and $\bar{2}$ have three outgoing edges, states 1 and 2 in M_1 each is split to three states, namely, $1'$, 1^*, 1^+ and $2'$, 2^*, 2^+. This implies that state 13 has to be split into three states. Other states are split accordingly. The resulting enlarged state machine is denoted by M_1'. Clearly, G_{M_1} is a deterministic graph. It is easily seen that M_1' realizes M_1 through a mapping φ which takes all states s^α($\alpha = *, +, '$ or 0; where $s^0 = s$) in S' to s in S. Also, every irredundant cyclic cover c on M is equal to $\varphi(C')$ for some irredundant cyclic cover C' on M' with the same period. Hence, the bound on periodicity of M_1' is also an upper bound on periodicity of M_1. We observe, however, that this bound is not the best possible.

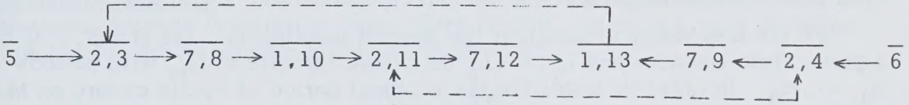


Figure 2. The cover graph of M_2

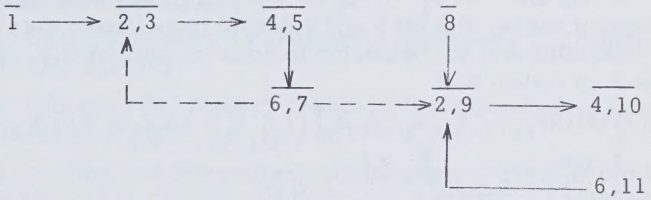


Figure 3. The cover graph of M_3

Consider now the third case. Let M be a given machine. For simplicity, we assume for the time being that there is only one vertex, say α^*, in G_M which has more than one outgoing edge. Let K denote the family of all simple circuits $\{C_i\}_{i\epsilon I}$ indexed by the set I in G_M. States of M are then split such that the cover graph of the resulting enlarged machine M' consists of a family of disjoint graphs of the following two types.

Each graph of the first type is isomorphic to a graph $G_1 = (V_1, \rho_1)$, where $V_1 = C_M$ and

$$\rho_1 = \{(\alpha, \beta) \mid [(\alpha, \beta) \in \rho_M \wedge \alpha \neq \alpha^*] \vee [\alpha = \alpha^* \wedge (\alpha^*, \beta) \in \rho_M \wedge$$
$$(\forall C_i \in K)[(\alpha^*, \beta) \notin C_i]]\}.$$

Each graph of the second type is isomorphic to a graph $G_2 = (V_2, \rho_2)$, where V_2 is a union of two sets V_{21} and V_{22} such that for each $C_i \in K$,

$$V_{21} = \{\alpha \mid \alpha \in C_i\}:$$
$$V_{22} = \{\alpha \in C_M \mid (\exists \alpha = \alpha_1, \cdots, \alpha_n \in C_M)[(\forall 1 \le i < n)[(\alpha_i, \alpha_{i+1}) \in \rho_M \wedge [\alpha_n \in C_i]\}.$$

And $\rho_2 = \{(\alpha, \beta) \in V_2^2 \mid (\alpha, \beta) \in \rho_M\}$.

We see that a graph of the first type is either a deterministic graph or a non-deterministic graph discussed in case 1, whereas graphs of the second type are either a deterministic graph or a non-deterministic graph satisfying the conditions of case 2.

As in case 1, it is seen that M_3' (transition table of M_3' can be easily constructed from G_{M_3}) realizes M_3, and that upper bounds for cyclic covers of M_3' are also the upper bounds for cyclic covers on M_3.

In case that M contains more than one vertex with multiple outgoing edges, then one simply picks a particular such vertex and proceeds with the state splitting mentioned above. The procedure is reapplied to components of the resulting cover graph which are of the same type as G_M. Therefore, case 3 also poses no problem.

We are now ready to consider the second possibility. Let $M = (S, \Sigma, \delta)$ be a given state machine and $L_1, \cdots, L_n$ be simple circuits of G_M with periods $m_1, \cdots, m_n$. In order to determine the maximal period of cyclic covers on M, it is necessary to know how many ways each vertex with multiple outgoing edges can be combined with other elements of G_M to form classes of some cyclic covers. To this end, let $\alpha_1, \cdots, \alpha_k$ be vertices of G_M each of which has multiple outgoing edges, and let E and F denote respectively sets of vertices in G_M belonging and not belonging to some circuits of G_M. We define a relation θ on F such that

$$\theta = \left\{(\alpha, \beta) \mid (\exists \gamma \in E)(\exists P_1, \cdots, P_n, q_1, \cdots, q_n)\left[[\sum_{i=1}^{n} P_i > 0 \wedge \sum_{i=1}^{n} q_i > 0] \wedge [\alpha \circ \rho_M^{\sum_{i=1}^{n} p_i m_i} = \beta \circ \rho_M^{\sum_{i=1}^{n} q_i m_i} = \gamma]\right]\right\}$$

It is seen that θ is a compatibility relation such that its corresponding cover C_θ has one class $T(\gamma)$ corresponding to each element $\gamma \in E$ (some $T(\gamma)$ may be empty though). Assume that M is connected. Let C be an irredundant cyclic cover on M such that the least and largest number of elements of E contained in classes of C are ℓ_1 and ℓ_2, respectively. We observe that the maximum number of classes which can be formed with each $\alpha_i (i=1, \cdots, k)$ is clearly

$$N_i = \max_{1 \le q \le |T(\alpha_i)|} \binom{|T(\alpha_i|}{q}$$ by Sperner's theorem [4], if $|T(\alpha_i)| \neq \emptyset$.

and is equal to 1 otherwise. Each such class can be combined with $t (\ell_1 \le t \le \ell_2)$ elements of $E - \{\alpha_i\}$. Thus, the maximum number of classes of C which contain the element α_i is $N_i \times_{\ell_1 \le t \le \ell_2} (|E_t| - 1)$. Since the average distance between such classes in C is $\left[\max_{1 \le i \le n} \min_{1 \le k \le \ell_1} \frac{m_i}{k}\right] = \left[\max\{\frac{m_1, \cdots, m_n}{\ell_1}\}\right]$ where $[\frac{P}{q}]$ denote the smallest integer greater than or equal to $\frac{P}{q}$, the period of C is clearly bounded by $\sum_{i=1}^{k}\left(N_i \times \max_{\ell_1 \leq r \leq \ell_2}\binom{|E|-1}{r} \times \left[\frac{\max(m_1^i, \dots, m_n^i)}{\ell_1}\right]\right)$ where m_j^i is the jth circuit containing α_i. The above result is summarized in the following theorem.

Theorem 10. Let M be a connected state machine and $\alpha_1, \dots, \alpha_k$ be vertices of G_M such that α_i is common to $n_i (n_i > 1)$ simple circuits $L_1^i, \dots, L_{n_i}^i$ with periods $m_1^i, \dots, m_{n_i}^i$. Let E, F, and $T(\alpha_i)$ be as defined in the preceding paragraph. If C is an irredundant cyclic cover on S such that the least and largest number of elements of E contained in classes of C are ℓ_1 and ℓ_2, respectively, then the period of C is bounded by

$$\sum_{i=1}^{k}\left\{\max_{1 \leq q \leq N(i)}\binom{N(i)}{q} \times \max_{\ell_1 - 1 \leq r \leq \ell_2 - 1}\binom{|E|-1}{r} \times \left[\frac{\max(m_1^i, \dots, m_n^i)}{\ell_1}\right]\right\}$$

where N(i) equals to $|T(\alpha_i)|$ if $T(\alpha_i) \neq \emptyset$, and equals to 1 otherwise.

Acknowledgements - The research reported here is supported in part by the National Science Foundation under grant GJ-786 and by the Joint Services Electronics Program under grant AFOSR 69-1792.

REFERENCES

1. J. Hartmanis and R.E. Stearns, Algebraic Structure Theory of Sequential Machines.
2. A. Gill, "Time-Varying Sequential Machines," J. Frankl. Inst. 276, (1963), 519-539.
3. A. Gill and J.R. Flexer, "Periodic Decomposition of Sequential Machines," JACM 14, No. 4 (1967), 666-676.
4. G.C. Rota, "A Generalization of Sperner's Theorem," Research Problem in J. Combinational Theory 2 (1967), 104-105.
5. S. Winograd, "Review of 'Time-Varying Sequential Machine' by Gill," IEEE EC-13, No. 6 (1964), 757.
6. R.T. Yeh, "On Relational Homomorphisms of Automata," Inf. Cont. 13 (1968), 140-155.

CLOSURE PROPERTY OF LINEAR SEQUENTIAL MACHINES

K. C. Wang
Washington State University
Pullman, Washington

ABSTRACT

This paper studies the variations of sequential machines. Two specific types of variations, the dual and the complement, of a given machine are considered. It investigates the relationship between the derived machines and the original machine and determines the class of machines that is closed and invariant under these operations. It is shown that the entire class of machines is closed under dual, and that the class of linear machines is closed and invariant under complement.

I. INTRODUCTION

In studying the realization of sequential machines, an interesting and important problem is to consider the variations of machines with known structure. Previous investigations on the homomorphic image of machines [1] and the automomorphism group of strongly connected machines [2] are typical examples along this line. This paper considers a similar problem but from an entirely different view point. It is motivated by the fact that a given machine can be used to derive a class of new machines if one modifies, say the next state function of the given machine according to some well defined rule. The question is then what are the derived machines? This problem is practically important in that it can be used to extend the realizability of machines and applied to fault diagnosis. This paper considers two special types of variations; namely the dual and the complement of a given machine. In particular, it investigates the relationship between the derived machines and the original machine, and determines the class of machines that is closed and invariant under dual and complement. The main result of this paper is to show that the entire class of linear machines is closed and invariant under complement.

II. PRELIMINARY

Let M = (S,I,f) be a state machine, where S is the set of states, I is the set of inputs and $f: SXI \rightarrow S$ is the next state function. Assume M is given in the form of a logic network consisting of m input lines and n delays. Then, all 2^n states and 2^m inputs of M are present, and f consists of a set of n Boolean functions of n state variables and m input variables.

Definition 1. The dual of a machine M is defined as

$$M_d = (S,I,f_d)$$

where f_d = the dual function of f.

Definition 2. The complement of M is defined as

$$\bar{M} = (S,I,\bar{f})$$

where $\bar{f}$ = the complement of f.

Definition 3. An isomorphism on M is an operation preserving 1-1 mapping on M. If the mapping is such that $S \rightarrow \bar{s}$ and $i \rightarrow \bar{i}$, it is called a shifting isomorphism.

Definition 4. An operation T acts on M is to modify the next state function f by a well defined rule. A machine M is invariant under T if the resulting machine is isomorphic to M. A class of machines is closed under T if the resulting machines belong to the same class.

III. STATEMENT OF PROBLEM

The specific problems to be considered in this paper are as follows. Given a machine M, determine its dual and its complement. Also, determine the class of machines that is closed and invariant under the operation of complement.

IV. RESULTS

First, we show the following:

Lemma 1. Every machine M is dual-invariant.

Proof. Let M be realized by a logic network. Then the same logic network also realizes the dual of M if negative logic is used. Under the negative logic realization, a state vector s becomes $\bar{s}$, and an input vector becomes $\bar{i}$. The transitions of M are obviously preserved. Hence M is shifting isomorphic to its dual.

It follows that the entire classes of all machines is closed and invariant under dual. However, it is easy to show that the complement of a machine M, in general, depends on the state functions, which in turn depends on the state assignment, of M. In fact, there are machines that do not have any complement-invariant realization. To see what kind of machines possess a shift-isomorphic realization, we need the following theorm.

Theorem 1. The necessary and sufficient condition for a machine M to have complement shift-isomorphic realization is that the states and the inputs can be partitioned into disjoint pairs (s,s^*) and (i,i^*) such that $f(s,i) = s'$ if and only if $f(s^*,i^*)= s'^*$.

Proof. Necessity: Assume Mand $\bar{M}$ are shifting isomorphic.

If $f(s,i) = s'$

then $\bar{f}(\bar{s},\bar{i}) = \bar{s}'$, for all s in S and i in I.

Since $\bar{f}(\bar{s},\bar{i}) = f_d(s,i)$, and by Lemma 1 we have

$f(\bar{s},\bar{i}) = \bar{s}'$.

Thus, $(s,\bar{s})$ and $(i,\bar{i})$ are the state and input pairs.

Sufficiency: Assume the states and inputs are partitioned into disjoint pairs. Divide the states (inputs) into two groups which separate the pairs. Assign arbitrarily a set of codes to the states (inputs) in one group; the codes of the remaining states (inputs) are simply the complement of their corresponding pairs. This assignment satisfies the requirement that if $f(s,i)= s'$, then $f(\bar{s},\bar{i}) = s'$ which is $\bar{f}(\bar{s},\bar{i}) = \bar{s}'$.

Following Theorem 1, it is easy to check whether a state table has a shift-isomorphic realization under complement. If it does, a valid state assignment can also be obtained. The algorithm is outlined below.

Algorithm:

Step 1. Check the table for pairs of symmetrical columns. If the states and inputs can be partitioned into disjoint pairs with symmetrical transitions, the table must consist entirely of pairs of symmetrical columns. This step, if successful, generates the input pairs and all possible state pairs implied by the symmetrical transitions.

Step 2. Check the state pairs for disjointness and consistency.

Step 3. If steps 1 and 2 are passed, the state assignment can be made in accordance with the pairs as pointed out in Theorem 1.

Example: Consider the following state table.

S \ I	i_1	i_2	i_3	i_4
s_1	s_1	s_3	s_1	s_2
s_2	s_3	s_1	s_2	s_1
s_3	s_3	s_2	s_4	s_2
s_4	s_2	s_3	s_2	s_4

Examining the columns of the table, we see that the two symmetrical pairs are (i_1,i_2) and (i_3,i_4). The possible state pairs are (s_1,s_2) and (s_3,s_4) which are consistent with the two input pairs. So we make the following assignment.

$s_1 = 00$ $\quad$ $s_2 = 11$
$s_3 = 10$ $\quad$ $s_4 = 01$
$i_1 = 00$ $\quad$ $i_2 = 11$
$i_3 = 10$ $\quad$ $i_4 = 01$

The next state functions are

$$Y_1 = \bar{x}_2 y_2 + x_2 \bar{y}_2 + x_1 x_2 \bar{y}_1 + \bar{x}_1 \bar{x}_2 y_1$$

$$Y_2 = \bar{x}_1 x_2 \bar{y}_2 + \bar{x}_1 \bar{y}_1 y_2 + x_1 y_1 \bar{y}_2$$

It is easy to verify that the transitions of M and $\bar{M}$ are shifting isomorphic.

We now consider the invariance property of linear machines. The objective is to show that the entire class of linear machines is closed and invariant under complement.

By definition [3], a machine M is linear if its state transitions are described by a matrix equation

$$s' = As + Bi \qquad (1)$$

where A and B are matrices over the field of 2 elements.

Lemma 2. In the state table of a linear machine M every pair of columns is symmetrical. Moreover, it is always possible to partition the inputs into pairs such that the implied state pairs are consistent.

Proof. The state table of M can be arranged in such a way that the first column (row) corresponds to the zero input (state). Denote the entries of the first column (row) by s_{p0}, $p=0,1,..,2^n-1$ (s_{0j}, $j=0,1,..,2^m-1$). Then the (p,j)th entry of the table is $(s_{p0}+s_{0j})$.

Consider any two columns i_j, i_k and two rows s_p, s_q. Assuming

$f(s_p,i_j)=f(s_q,i_k)$ we have

$$s_{p0}+s_{0j} = s_{q0}+s_{0k} \tag{2}$$

This implies that

$$s_{p0}+s_{0k} = s_{q0}+s_{0j} = s' \tag{3}$$

Since s' is precisely the next state of s_p under i_k and s_q under i_j, it follows that $f(s_p, i_k) = f(s_q, i_j)$ the two columns are, therefore, symmetrical. To show that there exists a valid pair partition of I and S, assume that i_0 has been paired arbitrarily with, say i_1 and one of the implied state pair is (s_p,s_q). Pick a new column i_j and let the (p,j)-th entry be $s'= (s_{p0}+s_{0j})$. We must find a column i_k with a leading entry s_{0k} such that

$$s_{0k}+s_{q0} = s' \tag{4}$$

or

$$\begin{aligned} s_{0k} &= (s_{p0}+s_{q0}) + s_{0j} \\ &= s_{01} + s_{0j}. \end{aligned} \tag{5}$$

Since the first row of the table form a subspace of s, s_{0k} is always present in the first row. Thus the column i_k can always be found. Continue this process we can generate a pair partition on I such that all the state pairs implied by (i_0,i_1) are consistent with the subsequent input pairs.

Combining Theorem 1 and Lemma 2, we have

Corollary 1. Every linear machine has a complement-invariant realization.

It is noted that this realization may not be linear. To maintain linearity for the shifting isomorphism, we need

$$\begin{aligned} As+Bi &= A\bar{s} + B\bar{i} \\ &\quad + As + Bi + A\alpha_S + B\alpha_I, \end{aligned} \tag{6}$$

where α denotes the column vector of all 1's. Eq.6 is equivalent to

$$A\alpha_S = B\alpha_1.$$

Here we have two cases:

Case 1. If the first row and the first column of M have an element $s \neq s_{00}$ in common, we simply apply a nonsingular linear transformation on S to change the predecessor of s to α_S. Then, change the corresponding input to α_I. These are always possible. The resulting codes satisfy $A\alpha_S = B\alpha_I$. M and $\bar{M}$ are shifting isomorphic in this case.

Case 2. If the only common element is s_{00}, we turn to the results of [4] and [5]. It was shown that if the characteristic polynomial $f(x)$ of A does not contain $(x-1)$ as a factor, i.e. $f(1) \neq 0$, then the state diagrams generated by

$$s' = As + Bi \qquad (7)$$

and

$$s' = As + Bi + \alpha_S \qquad (8)$$

are identical. Since this holds for every input i without complement, there is an isomorphism between the states in Eq. 7 and Eq. 8. The latter is precisely the state equation of $\bar{M}$. Hence M and $\bar{M}$ are isomorphic. In case $f(1) = 0$, it suffices to consider A to be the companion matrix of $(x-1)^n$. The isomorphism still holds as long as α_S is not a cyclic vector of A. When this is not the case we notice that, because A is nonsingular, the first column of M consists of the entire S. All the states in the first row of M must also be in the first column. This leads us back to Case 1 if M has two or more columns. If M has only one column, i.e. M is autonomous, and α_S is a cyclic vector of A, it is always possible to find a nonsingular P such that $P\alpha_S$ is not a cyclic vector. After this transformation, M and $\bar{M}$ are again isomorphic.

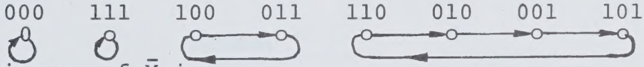
Example: Let M be an autonomous LSM described by

$$s' = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} s$$

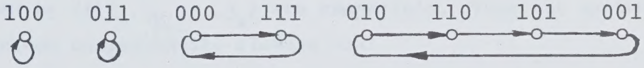
where A is the companion matrix of $(x-1)^3$. The state diagram of M is a cycle set (1,1,2,4). The vector $\alpha_S = 111$ is in the cycle of length 4; so α_S is a cyclic vector of A. The state diagram of $\bar{M}$ would be a cycle set (2(4)). Let

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Then $P\alpha_S = 101$, a noncyclic vector. After this transformation on S the state diagram of M becomes



The state diagram of $\bar{M}$ is



V CONCLUSION

In this paper, we have studied two specific types of machine variations; the dual and the complement. It is shown that every machine is dual-invariant, and that the class of linear machines is closed and invariant under complement. Further studies along this line would be to investigate machines subject to a subset of state and/or input variable changes.

REFERENCES

1. J. Hartmanis and R. E. Stearns, Algebraic Structure Theory of Sequential Machines, Prentice-Hall, 1966.
2. A. C. Fleck, "On the Automomorphism Group of an Automaton", J.ACM 12, pp. 566-569, Oct., 1965.
3. S. S. Yan and K. C. Wang, "Linearity of Sequential Machines", IEEE Trans. Computers, EC-15, pp. 337-354, June, 1966.
4. K. C. Wang, "Transition Graphs of Affine Transformation on Vector Spaces over Finite Fields", J. Frank Inst., Vol. 283, pp. 55-72, Jan., 1967.
5. A. Gill, "Graphs of Affine Transformations, with Applications to Sequential Circuits", Conf. Rec. 7th Symp. Switching and Automata Theory, pp. 127-135, 1966.

A METHOD FOR FINDING MAXIMUM COMPATIBILITY CLASSES FOR AN INCOMPLETELY SPECIFIED FSS

MOIEZ A. TAPIA
Georgia Institute of Technology, Atlanta, Georgia

ABSTRACT

A method for finding the maximum compatibility classes for a finite-state system with an incompletely specified transition table is presented. Unlike in the existing methods, the set of pairs of incompatible states is found and made use of. This set, with the aid of an algorithm given, leads to the generation of maximum compatibility classes.

SUMMARY

In order to find the set of pairs of incompatible states, we define a couple of terms. Two states s_i and s_j, $i \neq j$, are called 1-incompatible, if there exists at least one input $\underline{x} \in X$ such that output $\omega(\underline{x}, s_i) \neq$ output $\omega(\underline{x}, s_j)$. Two states s_i ans s_j, $i \neq j$, are called k-incompatible, $k > 1$, if they are not (k-1)-incompatible and there exists an input sequence, $\underline{x_k}$, of length k, such that the terminal output, at the end of sequence $\omega_T(\underline{x_k}, s_j) \neq \omega_T(\underline{x_k}, s_i)$. I_k, $k \geq 1$, denotes the set of all pairs of k-incompatible states. I_1 is found by comparing all possible pairs of rows in the transition table. In order to obtain I_{k-1}, consider a pair (s_i, s_j) in I_{k-1}. If s_i and s_j appear in any column, then the states corresponding to the rows in which they appear are k-incompatible, provided the latter pair is not in any I_p, $1 \leq p \leq k-1$. Repeating the procedure for all pairs in I_{k-1}, we obtain the set I_k. It can be shown that $I_k \equiv \phi$ for all $k > k_o$, $k_o \leq \frac{n(n-1)}{2}$. Taking union of all I_k's we obtain I, the set of all pairs of incompatible states. Write $I = \{i_1 | j_1, ---, j_{k_1}), \ldots, (i_r | t_1, ---, t_{k_r})\}$ such that $i_1 < i_2 < --- < i_r$, $1 \leq r \leq n-1$ and $(i | j_1, ---, j_k)$ denotes set of all states s_{j_m}, $| \leq m \leq k$, that are incompatible with s_i, $i < j_1 < --- <_{j_k}$. If $P = \{i_1, i_2, \ldots, i_k\}$ and $R = \{j_1, j_2, \ldots, j_p\}$, then $\frac{P}{R}$ denotes set of states denoted by integers in P but not in R. $\frac{P}{R} \triangleq \{s_i | i \in P, i \notin R\} \triangleq P$ "divided" by R.

THEOREM

Let an FSS be a system with n states such that all pairs of states except those denoted by $(i | j_1, j_2, \ldots, j_k)$ are compatible. Then the system has exactly two maximum compatibility classes viz.

$$\frac{(i_1, i_2, \ldots i_n)}{(i)} \quad \text{and} \quad \frac{(i_1, i_2, \ldots, i_n)}{(j_1, j_2, \ldots, j_k)}$$

ALGORITHM

Step 0: Form $M_o = \{(i_1, i_2, \ldots, i_n)\}$ i.e., a maximum compatibility class consisting of all states in the system.

Step i ($i \geq 1$): If $(i_i | j_1, j_2, \ldots, j_i)$ is the $i^{\underline{th}}$, term in I, then "divide" each maximum compatibility class in M_{i-1} by (i_i) and $(j_1, j_2, \ldots, j_i)$. A compatibility class in M_{i-1}, in general, will give rise to two compatibility classes in M_i. Strike out classes which are proper subsets of classes in M_i. If any compatibility classes are repeated, stricke out all but one of them.

Step (r+1): Stop. All the classes obtained in M_r are maximum compatibility classes.

SOME NEW RESULTS CONCERNING SEPARATION THEORY OF GRAPHS

DONOVAN C. PLISCH
Bell Telephone Laboratories
Naperville, Illinois

DAVID P. BROWN
University of Wisconsin
Madison, Wisconsin

Abstract: Two fundamental problems which occur in the separation theory of graphs may be stated as follows: Given a graph G with two identified disjoint vertex sets A and B.

(1) Separate A and B by removing the minimum number of edges from G.

(2) Separate A and B by removing the minimum number of vertices from G.

The authors describe new and simplified methods for solving both of these problems. Applications for these results are in the areas of integrated circuits packaging, studies of k-connectivity of graphs, communications networks and transportation networks. Computer programs have been written based on the theory.

I. INTRODUCTION

Two fundamental theorems from separation theory of graphs found in the literature [1] may be stated as follows:

Given a graph G with two identified disjoint vertex sets A and B where the edges of G have no orientation.†

Theorem 1: The minimum number of edges in a set of edges E_j which if removed from G will place A and B in separate subgraphs is equal to the maximum number of edge disjoint paths which can be labeled between A and B.‡

Theorem 2: The minimum number of vertices in a set of vertices V_k which if removed from G will place A and B in separate subgraphs is equal to the maximum number of vertex disjoint paths which can be labeled between A and B.

If the set E_j contains the minimum number of edges it is known as a minimum edgecut $E_j{*}$. If the set V_k contains the minimum number of vertices it is known as a minimum vertex cut $V_k{*}$.

Theorem 1 is illustrated in Figure 1 for which a graph G is shown with two disjoint vertex sets A and B. There are three edge disjoint labeled paths between A and B, and three edges in the minimum edgecuts separating A and B. Theorem 2 is illustrated in Figure 2 for the same graph G as used in Figure 1. There are two vertex disjoint labeled paths between

---

† The edges of G initially have no orientation. However, orientation will be added to certain edges in what follows.

‡ A path is a connected subgraph of G which has every vertex of degree 2 except for two vertices of degree 1.

A and B, and two vertices in the minimum vertex cuts separating A and B.

Given G with two identified disjoint vertex sets A and B, a fundamental problem is to determine all of the minimum edgecuts $\{E_j^*\}$ separating A and B. Another similar problem is to determine all of the minimum vertex cuts $\{V_k^*\}$ separating A and B. In this paper new and efficient methods are given for solving both of these problems. An outline of the approach used for both problems given G, A, B is as follows:

(1) Label edge (vertex) disjoint paths between A and B using a shortest path algorithm. (Use of the shortest path algorithm is not absolutely necessary but usually speeds up the solution.)

(2) Maximize the number of labeled paths by rearranging and adding labels to edges of G.

(3) Select minimum edgecuts (vertex cuts) from the edges (vertices) of the labeled paths of G.

Applications for the results are found in the areas of integrated circuits packaging, studies of k-connectivity of graphs, communications networks and transportation networks.

II. DETERMINING $\{E_j^*\}$

The problem of determining $\{E_j^*\}$ is related to a problem investigated by Ford and Fulkerson [2]. The method of solution given here is considerably different than in [2]. Several useful definitions are given below which apply to a graph G for which two disjoint vertex sets A and B have been identified.

An α-labeled path $P_{\alpha_r}(B,A)$ is a path from B to A which has every edge labeled with one and only one symbol α_r where α_r is the same for every edge of the path. Every edge of $P_{\alpha_r}(B,A)$ is oriented from a vertex $b \in B$ to a vertex $a \in A$. No degree 2 vertex of $P_{\alpha_r}(B,A)$ is contained in the sets A or B.

The graph of Figure 3 has two α-labeled paths. (Note that α-labeled paths are edge disjoint by definition.)

A relabeling path R(A,B) is a path from a vertex $a \in A$ to a vertex $b \in B$ with the property that if an edge of R(A,B) is an oriented edge then it is directed away from $a \in A$. No degree 2 vertex of R(A,B) is a vertex of the sets A or B.

The vertices of a relabeling path for Figure 3 are 1, 3, 6, 4, 5, 7, 9. Note that a path of unlabeled edges from A to B also meets the definition of a relabeling path.

Two theorems concerning relabeling paths are now presented which yield a method for determining a maximum number of α-labeled paths for G, A, B.

Theorem 3: Given a graph G with two identified disjoint vertex sets A and B for which there are $k \geq 0$ α-labeled paths. If a relabeling path R(A,B) exists for G, A, B then by changing and adding α-labels to the edges of G the number of α-labeled paths can be increased to k + 1.

Proof: (The algorithmic method for maximizing the number of α-labeled paths is based on the proof of this theorem.) Assume that R(A,B) has edges from k different α-labeled paths. If k = 0 go to step 2; otherwise start with step (1).

(1) It follows directly from the definitions of an α-labeled path and a relabeling path R(A,B) that the first j edges of R(A,B) are unlabeled where $j \geq 1$. Consider the vertex v which is the first degree 2 vertex of R(A,B) that is incident to both an unlabeled edge and an edge of an α-labeled path. There must exist two paths between v and the vertex set A:

(a) $P_{\alpha_r}(v,A)$ which is a subpath of an α-labeled path with edge labeling symbol α_r.

(b) P(v,A) which is a subpath of R(A,B) consisting of unlabeled edges.

Remove the edge labels and the edge orientation from all edges of $P_{\alpha_r}(v,A)$. Add the label α_r of (a) to all edges of P(v,A) and change the edges of P(v,A) to directed edges (directed towards $a \in A$).

After the above relabeling there will exist a new relabeling path R'(A,B) which has fewer edges in common with α-labeled paths. If R'(A,B) has no edges in common with α-labeled paths then go to step (2). Otherwise repeat step (1) for R'(A,B).

(2) By repeated application of step (1) a relabeling path R"(A,B) exists with no labeled edges; that is, R"(A,B) has no edges in common with any α-labeled path. The path R"(A,B) can be changed to an α-labeled path by changing the edges to directed edges (directed away from $b \in B$ where b is an end vertex of R"(A,B)) with a symbol α_s where α_s has not previously been used. This increases the total number of α-labeled paths for G, A, B to k + 1 which proves the theorem.

The application of Theorem 3 and its proof is illustrated in the graph of Figure 3 for which a relabeling path exists whose vertices are 1, 3, 6, 4, 5, 7, 9. Vertex 6 is vertex v, $P_{\alpha_2}(6,2) = P_{\alpha_r}(v,A)$ and P(v,A) is the path whose vertices are 6, 3, 1. If step 1 of the proof is applied to Figure 3 then a path of unlabeled edges from A to B is created whose vertices are 2, 4, 5, 7, 9. A new α-labeled path can then be added as in Figure 4. That is, the number of α-labeled paths has been increased from two to three.

Theorem 4: Given a graph G with two identified disjoint vertex sets A and B, and k α-labeled paths. If and only if no relabeling path exists for G, A, B then k is maximum.

A proof of Theorem 4 is given in Reference [4]. Theorem 4 is illustrated in Figure 4 for which no relabeling path exists.

The number of α-labeled paths is three which is also the maximum number by Theorem 4.

The above theorems suggest the following procedure for determining the maximum number of α-labeled paths for G, A, B where the edges of G are initially unlabeled:

(1) Determine a relabeling path; initially chosen relabeling paths will consist entirely of unlabeled edges. (The amount of label changing necessary to determine a maximum number of α-labeled paths is usually minimized by chosing the shortest relabeling path first.)

(2) Change labels as in the proof of Theorem 3 so as to increase the number of α-labeled paths by one. Repeat until no additional relabeling paths exist.

Edgecut Selection

Two properties of G, A, B with a maximum number of α-labeled paths follow directly from Theorem 1:

(1) A minimum edgecut has one and only one edge from every α-labeled path.

(2) If an edge of G is not an edge of an α-labeled path then it can be in no minimum edgecut.

Now consider the following rules for assigning vertices of G to subgraphs $\{G_\ell\}$ where V_ℓ is the vertex set of G_ℓ:

(a) Assign A to V_1 and B to V_2.

(b) If there is a path of unlabeled edges between two vertices v_1 and v_2 then v_1 and v_2 are assigned to some one V_ℓ for G_ℓ.

(c) If a vertex v of G has no unlabeled edge incident to it and $v \notin A, B$ then v is defined to be some V_ℓ for G_ℓ.

For Figure 4, $\{V_\ell\} = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9, 10\}\}$. A consideration of the above properties and assignment rules leads to the conclusion that an edge e of G can be an edge of a minimum edgecut only if e has vertices in two different G_ℓ.

Now select a set of edges $\{e_p\}$ which contains one and only one edge from each α-labeled path. Also, every $e_p \varepsilon \{e_p\}$ has vertices in two different G_ℓ. If $\{e_p\}$ separates G into two subgraphs with vertex sets X, Y where $A \in X$, $B \in Y$ then $\{e_p\}$ is by definition a minimum edgecut $E_j{}^*$. (For Figure 4, $\{e_1, e_2, e_3\}$ is a minimum edgecut but $\{e_1, e_2, e_4\}$ is not.) It is easy to show that all sets of edges which satisfy the conditions of this paragraph is identically the same as $\{E_j{}^*\}$. (Reference [4]).

III. DETERMINING $\{V_k^*\}$

The set of all minimum vertex cuts $\{V_k^*\}$ can be determined in a manner parallel to the way in which $\{E_j^*\}$ is determined. In the theory of Section II make the following replacements:

γ-labeled path → γ-labeled path
vertex cut → edgecut
removal of vertices → removal of edges

In addition add the following requirements to the definitions:

(1) No two γ-labeled paths may have a degree 2 vertex in common.

(2) If two unlabeled edges of a relabeling path have a vertex v in common then v cannot be a vertex of a γ-labeled path.

When the above changes have been accomplished, the theory of Section II can be used to determine a maximum number of vertex disjoint γ-labeled paths. However, it is inconvenient (although not impossible) to apply this parallelism to the problem of determining $\{V_k^*\}$. Figure 5 shows a graph which has two γ-labeled paths and a relabeling path whose vertices are 1, 3, 4, 5, 7, 6, 10.

A method of selecting the set of all minimum vertex cuts $\{V_k^*\}$ will now be illustrated with an example. The details of the method are given in Reference [4].

In Figure 6 a graph G is represented for which a maximum number of γ-labeled paths exist. A minimum vertex cut V_k^* for G would have one and only one degree 2 vertex from each γ-labeled path. A partial vertex cut is a subset of a vertex cut. Consider the following sets of vertices from G: {3, 4}, {3, 8}, {7, 4}, {7, 8}. The vertex set {3, 8} cannot be a partial vertex cut because the path from A to B whose vertices are 1, 4, 7, 10 would exist even if {3, 8} and a vertex from P_{γ_3} were removed from G. For a similar reason {7, 8} cannot be a partial vertex cut but no such path from A to B can be found for {3, 4} or {7, 4}.

An algorithm which determines $\{V_k^*\}$ by finding all sets of m + 1 vertices - which are possible partial vertex cuts for a set of m + 1 γ-labeled paths - from all possible partial vertex cuts for a set of m γ-labeled paths is not difficult to formulate. This simple procedure has proven to be effective on large graphs if the number of minimum vertex cuts is not excessive.

IV. OUTLINE OF PROGRAMMING METHODS

A computer program has been written based on the material presented in Section II for finding $\{E_j^*\}$. An intermediate output of this program is a maximum set of edge disjoint paths between A and B. This program will determine all minimum edgecuts $\{E_j*\}$ for a large graph with the only restriction that the number of minimum edgecuts cannot be excessive. A similar program has been written which determines $\{V_k*\}$ and a maximum number of vertex disjoint paths between A and B. A printout and detailed description of both programs is given in Reference [4].

An extremely useful way of representing a graph G using one-dimensional arrays used in both programs is illustrated in Figure 7. A graph G is shown whose vertex incidence patterns are stored in the arrays LOC1, LOC2 and EDGE where the entries in LOC1 and LOC2 refer to locations in EDGE. Consider vertex 2 of G which is of degree 5. For the arrays LOC1 and LOC2, LOC1(2) = 4 and LOC2(2) = 8. Locations 4 through 8 in EDGE contain the list of vertices 1, 1, 3, 4, 4 which are the vertices that have edges in common with vertex 2 of G. Note that vertices 1 and 4 appear twice in the list since they have two parallel edges in common with vertex 2.

The method illustrated in Figure 7 for representing a graph in one-dimensional arrays has proved to be an essential element of the programming effort. It is extremely efficient in terms of memory locations necessary. For a graph with n_e edges and n_v vertices only $2n_e + 2n_v$ memory locations are required in the arrays LOC1, LOC2, and EDGE. Also computational efficiency is excellent.

An additional advantage of representing a graph using the one-dimensional arrays of Figure 7 became apparent only after the first program had been written: Most of the algorithms require a great deal of extra labels on the vertices and edges of G. Consider for example G of Figure 7 which has an α-labeled path from vertex 1 to vertex 4, $P_{\alpha_3}(1,4)$. This fact may be easily noted in the program by using another one-dimensional array αLABEL which is the same size as EDGE. An edge of P_{α_3} is directed from vertex 2 to vertex 4. This is noted in αLABEL of Figure 7 by setting αLABEL(7) = +3 and αLABEL(12) = -3. Note that αLABEL may be addressed in identically the same way as EDGE. As an example of vertex labeling consider the vertex labels 16, 17, 18, 22 in parentheses beside the vertices of G in Figure 7. These labels are placed in the array VLABEL which is the same size as LOC1 and LOC2.

It is obvious from the example that it is a trivial matter to add as many vertex labels and edge labels to G as desired using any commonly available programming language such as Fortran which allows subscripted variables. Manipulation of these labels is also easy since the subscripting is identical for all arrays which contain vertex labels, and the subscripting is also identical for all arrays which contain edge labels.

References:

1. O. Ore, Theory of Graphs, American Mathematical Society Colloquium Publication, 1962.
2. L. R. Ford and D. R. Fulkerson, Flows in Networks, Princeton University Press, 1962.
3. I. T. Frisch, "An Algorithm for Vertex Pair Connectivity," Internat'l J. Control, Vol. 6, pp. 579-583, 1967.
4. D. C. Plisch, New Results Concerning Separation Theory of Graphs, University of Wisconsin PHD Thesis, 1970.

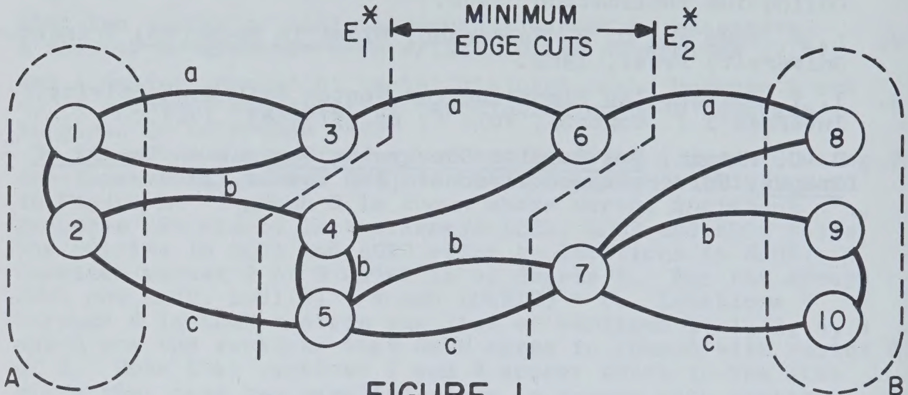

E_1^*
MINIMUM
EDGE CUTS
E_2^*
a
a
a
b
b
b
b
c
c
c
1
2
3
4
5
6
7
8
9
10
A
B

FIGURE 1

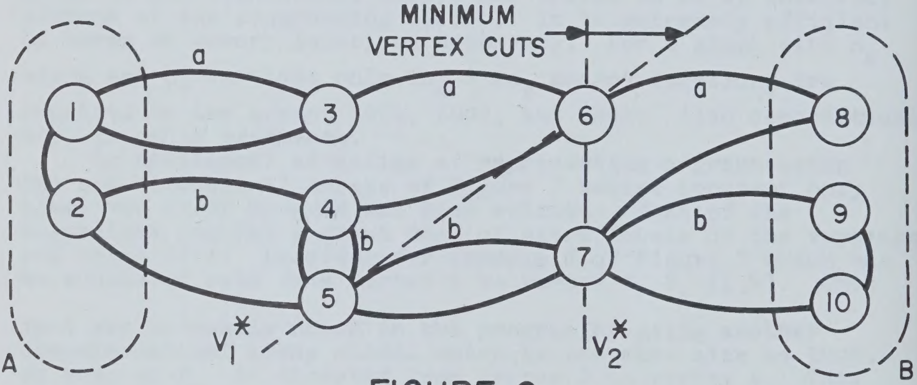

MINIMUM
VERTEX CUTS
a
a
a
b
b
b
b
1
2
3
4
5
6
7
8
9
10
V_1^*
V_2^*
A
B

FIGURE 2

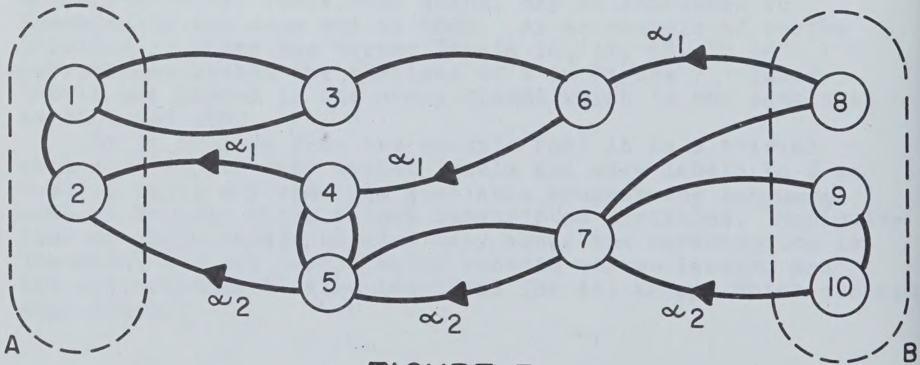

α_1
α_1
α_1
α_2
α_2
α_2
1
2
3
4
5
6
7
8
9
10
A
B

FIGURE 3

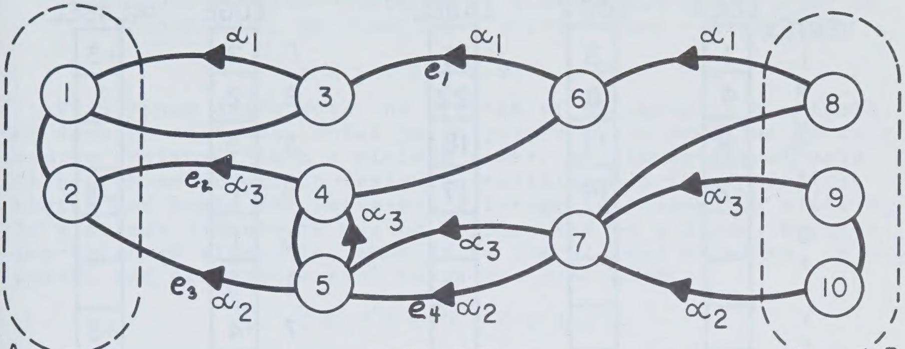

α1
α1
α1
e1
1
3
6
8
2
4
9
e2 α3
α3
α3
α3
5
7
10
e3 α2
e4 α2
α2
A
B

FIGURE 4

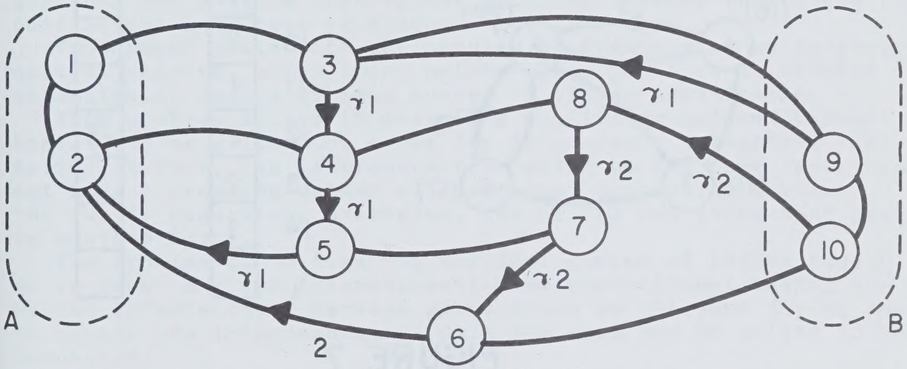

1
3
8
γ1
γ1
2
4
γ2
γ2
9
γ1
5
7
10
γ1
γ2
6
2
A
B

FIGURE 5

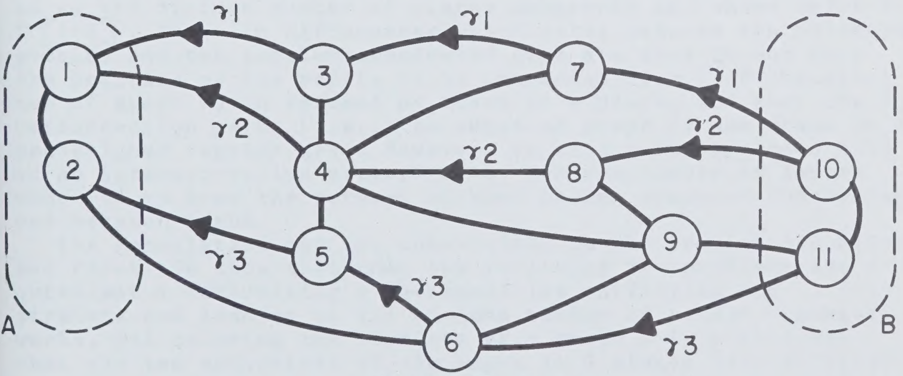
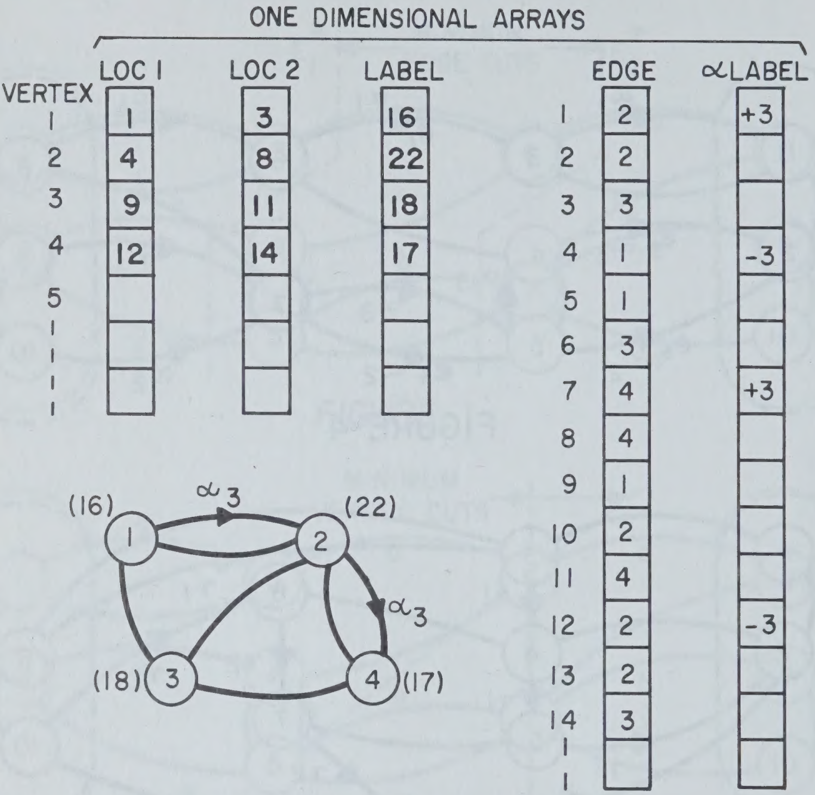

γ1
γ1
1
3
7
γ1
γ2
γ2
γ2
2
4
8
10
9
γ3
5
11
γ3
6
γ3
A
B

FIGURE 6


ONE DIMENSIONAL ARRAYS
(16)
1
α_3
(22)
2
α_3
(18)
3
4
(17)

VERTEX	LOC 1	LOC 2	LABEL
1	1	3	16
2	4	8	22
3	9	11	18
4	12	14	17
5			

	EDGE	αLABEL
1	2	+3
2	2	
3	3	
4	1	-3
5	1	
6	3	
7	4	+3
8	4	
9	1	
10	2	
11	4	
12	2	-3
13	2	
14	3	

FIGURE 7

ON THE PLANAR DECOMPOSITION OF INTERCONNECTION NETWORKS FOR COMPUTER-AIDED ELECTRONIC DESIGN

GIAMPIO BRACCHI
Istituto di Elettrotecnica ed Elettronica-Politecnico di Milano-P.za L. da Vinci,32 - 20133 Milano - Italy

ABSTRACT

This paper deals with the problem of decomposing a network of interconnections, which joins points whose position is assumed as assigned, into a minimum number of planar **subnetworks**. This problem arises in designing multilayer printed circuits boards and masks for large-scale integrated circuits, since only a planar subnetwork may be implemented on a layer. A computer-oriented algorithm, leading to the minimal solution, is proposed, and its theoretical bases are discussed.

1. INTRODUCTION

The increasing complexity of electronic systems has produced the need for computer-assisted design procedures, and thus for computer-oriented design methods. Among these methods, some algorithms for solving topological problems arising in printed circuit layout have been presented.

This paper deals with the problem of decomposing an interconnection network, which joins points whose position is assumed as assigned, into a minimum number of planar subnetworks.

This problem arises in designing multilayer printed circuit boards and metallization masks for large scale integrated circuits. In fact, two interconnection paths on the same layer cannot have a crossing unless an electrical connection between the two is requested; otherwise, one of the two paths must pass to another layer.

The problem is to find the smallest number of layers required to draw a given interconnection network without having unwanted interferences between paths drawn on the same layer, and to choose one interconnection path for each set of points to be connected.

A similar problem, which is known as thickness problem, exists in graph theory; the thickness $\theta(\Gamma)$ of a graph Γ is defined as the minimum number of planar subgraphs of Γ whose union is Γ [1-2]. The main differences which exist between the thickness problem and the problem considered here are that in our case the position of the points to be connected (i.e., of the vertices of graph Γ) is assumed as fixed in a plane, and that the interconnection paths (i.e., the edges of graph Γ) are drawn on a preassigned regular grid. However, we will call thickness $\theta(N)$ of an interconnection network N the minimum number of layers required to draw the network without having unwanted interferences between paths.

The formulation and the constraints to the problem are discussed first. We show then that the following two problems are equivalent : determining a decomposition satisfying the constraints and leading to the minimum number of planar **subnetworks**, and coloring the vertices of a graph G in such a way that the two end-points of the edges in G always have different colors, and the number of colors is minimum. A computer-oriented algorithm, based upon this equivalence, is proposed for find ing the thickness of the interconnection network, and for determining a minimal decomposition.

II. PROBLEM FORMULATION AND CONSTRAINTS

A printed circuit board or a slice of integrated circuits is considered as made up of cells on a regular pre-defined x-y grid GR; the points to be connected and the interconnection paths are restricted to lay between adjacent lines of the grid.

Let $M = \{m_1, m_2, m_e, \dots, m_n\}$ be the set of all the points to be connected by the interconnection network; these points will be called terminals.

Definition 1 - A subset $S_k = \{m_a, \dots m_q, \dots m_t\}$ of M, including points which must be joined together by an interconnection path, will be said to be a *common set* of points.

It is obvious that $M = \bigcup_{k=1}^{\bar{k}} S_k$ where $\bar{k}$ is the number of all possible common sets.

A finite number of possible interconnection paths, drawn on the grid GR, is assumed as preassigned for connecting the points of a common set.

Definition 2 - The set $\mathcal{R}_k = \{R_k^1, R_k^2 \dots R_k^f, \dots R_k^{s_k}\}$ including all the possible paths R_k^f $(f=1,\dots,s_k)$ which connect the points of a common set S_k will be said to be a *redundant path set*.

The members of $\mathcal{R}_k$ could correspond, for instance, to the possible minimal lenght interconnection paths, drawn on the preassigned grid, which connect the points of S_k. An algorithm for determining such paths is illustrated in reference [3].

In Fig. 1 a sample board is shown. The set of all points to be connected is $M = \{m_1, m_2, m_3, m_4, m_5, m_6\}$. Three common sets are assigned; they are: $S_1 = \{m_1, m_2, m_3\}$; $S_2 = \{m_4, m_5\}$; $S_3 = \{m_3, m_6\}$. Correspondingly, three redundant path sets are given; these are: $\mathcal{R}_1 = \{R_1^1, R_1^2\}$; $\mathcal{R}_2 = \{R_2^1, R_2^2\}$, and $\mathcal{R}_3 = \{R_3^1\}$. Interconnection paths R_1^1 and R_1^2, for example, correspond to minimal lenght path for the common set S_1.

In printed circuit and integrated circuit layout, through connections from one layer to another are allowed [4]. In this paper we will assume that the points of a path where a through hole may be introduced are preassigned; such points will be called *hole points*. For instance, one could impose that hole points may correspond to, and only to, points of a path where an horizontal segment and a vertical segment which belong to the path join together; such assumption will be made for the sample board shown in Fig. 1. Hole points may be considered as new terminal points to be added to set M. Actually, at a hole point we may introduce a through hole which separates a path into two new subpaths, each having the hole point as an end-point. When all the pre-assigned hole points are introduced on a path R_k^f, it is separeted into v_k^f new paths $R_k^{f,j}$; such new paths are called *elementary paths*, and they are said to belong to path R_k^f. We have $R_k^f = \bigcup_{j=1}^{v_k^f} R_k^{f,j}$.

From the interconnection paths shown in Fig. 1, introducing the hole points m_1^h and m_2^h in path R_1^1, m_1^h and m_3^h in path R_1^2,

m_4^h in path R_2^1 and m_5^h in path R_2^2, the elementary paths **shown** in Fig. 2 are obtained. In this figure, squares indicate terminals, and **circles** indicate hole points. As it can be seen, 11 elementary paths are obtained. We have, for example, $R_1^1 = (1 \cup 2 \cup 3 \cup 4)$. The elementary paths which are common to all the paths of a redundant path set are considered only once; for example, elementary paths 1 and 2 in Fig. 2, which are common to the two paths R_2^1 and R_2^2 (see Fig. 1) of the redundant path set $\mathcal{R}_1$, are considered only once.

In printed circuit and integrated circuit layout, technological or electrical reasons may impose the constraints that a certain path be implemented entirely on a single layer, or a pair of paths be implemented on the same layer, or two paths lay on different layers. Such constraints will be considered in the decomposition algorithm. We can introduce the following Definitions:

Definition 3 - A path is said to be uniplanar if it must be entirely drawn on a given layer.

Definition 4 - Two paths are said to be joined if they are uniplanar and they must be implemented on the same layer.

Definition 5 - Two paths A and B are said to be separated if any part of A and any part of B cannot be implemented on the same layer.

If we select a path from each redundant path set $\mathcal{R}_k$, we obtain an interconnection network N_i. Let s_k be the cardinality of set $\mathcal{R}_k$; then the number of possible interconnection networks is $u = \Pi_{k=1}^{\bar{k}} s_k$. These networks constitute a set $\mathcal{N} = \{N_1, N_2, \ldots, N_i, \ldots, N_u\}$. Our problem is to determine a network having the minimum thickness $\theta_{min} = \min_{i=1,u} \theta(N_i)$. Such minimum thickness will be denoted by $\theta_{min}(\mathcal{N})$.

In the next section the problem of determining $\theta_{min}(\mathcal{N})$ will be shown to be equivalent to the problem of coloring the vertices of a particular graph.

III. THE EQUIVALENT GRAPH

In our problem two paths interfere if they have a crossing or if they occupy at least one same cell.

Definition 6 - Two paths A and B, belonging to different redundant path sets, are said to be incompatible if they interfere in at least one point which is not a terminal point.

Two paths A and B are said to be compatible if they are not incompatible. The compatibility relation is both reflexive and symmetric but it is not transitive; thus, it is similar to the compatibility relation of switching theory [5-6] ; we may then introduce some well-known concepts of switching theory:

Definition 7 - A compatibility class C_h is a set of paths such that all the members of C_h are pairwise compatible.

Definition 8 - A compatibility class is maximal if it is not a proper subset of any other compatibility class.

The incompatibility relation on the set of paths can be represented by means of a graph G, which is called the incompatibility graph; each vertex in G is associated with a path; two vertices in G are joined by an incompatibility edge when the corresponding paths are incompatible. Since a path R_n^f is subdivided by its hole pojnts into v_k^f elementary paths $R_k^{f,j}$, in the incompatibility graph each elementary path is associated with a vertex. The v_k^f vertices which are associated with the elementary paths belonging to the same path R_k^f are joined together by edges of a second kind, which are called path edges, to form a complete subgraph.

Moreover, since in an interconnection network N_i only one path must be selected in each redundant path set, two vertices, associated with two elementary paths $R_k^{e,g}$ and $R_k^{o,h}$ belonging respectively to a path R_k^e and to a path R_k^o which are member of the same redundant path set $\mathcal{R}_k$, are joined by an edge of a third kind, which is called redundancy edge. For each pair $\{R_k^e, R_k^o\}$ these edges constitute a bipartite subgraph.

For the network shown in Figures 1 and 2, the incompatibility graph is shown in Fig. 3. Vertices 7 and 2, for example, are joined by an incompatibility edge, because simple paths 7 and 2 have a crossing; vertices 4 and 11 are joined by an incompatibility edge, because paths 4 and 11 occupy two same cells of the grid; vertices 7 and 8 are joined by a path edge, because simple paths 7 and 8 belong to the same path R_2^1; vertices 7 and 9 are joined by a redundancy edge, because paths R_2^1 and R_2^2 are members of the same redundant path set $\mathcal{R}_2$.

As it can be seen, the incompatibility graph G is an undirected graph with single edges and no loops, and it has edges of three different kinds.

Each network $N_i \in \mathcal{N}$ corresponds to a set of vertices of G; this set is denoted itself by N_i. In this set, no two vertices are joined by a redundancy edge, and subsets of vertices are joined by path edges to constitute complete subnetworks. The problem of determining the thickness $\theta(N_i)$ of network N_i is then equivalent to the problem of decomposing the vertices of N_i into K mutually disjoint classes, so that no two vertices in each class are joined by an incompatibility edge, and the number K of classes is minimum. If we consider each class to represent a color, then each vertex is colored in such a manner that the end-points of an incompatibility edge always have different colors. The thickness of network N_i is then equal to the chromatic number $K(G_i)$ (i.e., to the smallest number of classes in any coloration [7-8]) of the graph G_i having as vertices the members of set N_i, and as edges the incompatibility edges joining pairs of these vertices. Thus, the problem of determining among the u networks $N_i \in \mathcal{N}$ the one(s) having the minimum thickness is equivalent to the problem of finding, among the u graphs G_i, the one(s) having the minimum chromatic number $K(G_i)$.

Let us now discuss how the problem constraints are reflected in the incompatibility graph.

When a path R_k^f is uniplanar, (see Definition 3), no through holes are allowed; thus, no hole points are introduced, and only one vertex corresponds to the path. When two paths are joined (see Definition 4) they are necessarily uniplanar; the junction relation is reflexive, symmetric and transitive. Thus, it is an equivalence relation [6] . Hence, the junction relation divides the set of vertices of graph G into disjoint subsets, which will be called junction classes, so that the vertices in each class are pairwise joined, and no vertex in one class is joined to any vertex in any other class. Some well-known algorithms exist for determining equivalence classes, and thus junction classes, in graph G; see, for example, references [6] and [9].

From graph G a new graph G^* can be derived, whose vertices $J_1, J_2 \dots J_e$ correspond each to a junction class. Two vertices J_a and J_b in G^* are joined by an incompatibility edge when in G there exist two vertices, belonging respectively to junction classes J_a and J_b, which are joined by an incompatibility edge. Obviously, the order, i.e., the number of vertices, of G^* is less than or equal to the order of G. We can state the following :

Theorem 1 - The chromatic number $K(G)$ of G is equal to the chromatic number $K(G^*)$ of G^*, and from a coloring of G^* a coloring of G can always be derived.

Proof - Suppose that a coloration of G^* with K colours is obtained. If every vertex in G belonging to a junction class J_h is given the same color of vertex J_h in G^*, a coloration of G with K colors is obtained. In fact, vertices in G which are joined by an incompatibility edge have different colors. Thus we have : $K(G^*) \geq K(G)$. Suppose now that a coloration of G satisfying the constraints is obtained; vertices belonging to a junction class J_h are necessarily given the same color. If this color is given also to vertex J_h in G^* for every h, a coloring of G^* is obtained, since vertices in G^* which are joined by an incompatibility edge have different colors. Thus we have $K(G) \geq K(G^*)$. We conclude that $K(G^*) = K(G)$.

Q.E.D.

From Theorem 1 it follows that, in our problem the junction relations can be taken into account by considering graph G^* . However, since for each common set of points a redundant set of paths is generally given, the junction relation exists between pairs of redundant path sets and we have junction classes of redundant path sets, say $\mathcal{J}_h = \{\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_g\}$.

In any network $N_i \in \mathcal{N}$, one path is selected from each redundant path set belonging the $\mathcal{J}_h$. Such paths are pairwise joined; they constitute a junction set J_h^f; by Theorem 1 in graph G^* they are associated to a same vertex J_h^f . Since the number of all possible junction sets J_h^f is $r = (s_1 \cdot s_2 \cdot \dots \cdot s_g)$, where s_p is the cardinality of $\mathcal{R}_p$, in graph G^* these sets correspond to r vertices J_h^f, connected together by redundancy edges to form a complete subnetwork; thus, the junction sets J_h^f, for $f = 1, \dots r$, may be considered as the members of a new redundant path set.

Let us now consider separated paths (see Definition 5). Also the separation relation generally exists between a pair $(\mathcal{R}_g, \mathcal{R}_h)$ of redundant path sets, and it can be easily taken into account in graph G joining every pair of vertices associated to paths R_g^p and R_h^q by means of an incompatibility edge $(p = 1, \dots s_g ;\ q = 1, \dots s_h)$. If R_g^p and R_h^q are divided by hole points into elementary paths, every vertex $R_g^{p,c}$ $(c = 1, \dots v_g^p)$ is joined by an incompatibility edge with every vertex $R_h^{q,e}$ $(e = 1, \dots v_h^q)$; hence a bipartite subgraph is introduced in G.

The graph G' which is obtained by G taking into account the junction and the separation relations is said modified graph.

For instance suppose that for the network shown in Fig. 1 redundant path sets $\mathcal{R}_2 = \{ R_2^1, R_2^2 \}$ and $\mathcal{R}_3 = \{R_3^1\}$ are joined and thus uniplanar. A junction class of redundant path sets $J_1 = \{\mathcal{R}_2, \mathcal{R}_3\}$ will exist. In our case $r = (s_2 \cdot s_3) = 2.1=2$; thus, two junction sets of paths exist; these are $J_1^1 = (R_2^1, R_3^1)$ and $J_1^2 = (R_2^2, R_3^1)$, and they will correspond to two new vertices 12 and 13, joined by a redundancy edge, in the modified graph G'. This graph is shown in Fig. 4(a).

Let us now examine the effects of the introduction of the hole points on the thickness $\theta(N_i)$ of a network N_i. If a hole point is introduced on a path A belonging to N_i, A is separated into p parts ($2 \leqslant p \leqslant 4$), and thus into p new paths (A_1, ..., A_p). In the incompatibility subgraph G_i, corresponding to N_i, vertex A will be substituted by p vertices $A_1 \dots A_p$; the incompatibility edges which are incident in any of these new vertices are a subset of the incompatibility edges which are incident in vertex A in graph G_i. The following Theorem holds :

Theorem 2 - If a vertex A in graph G_i is substituted by p new vertices $A_1, \dots A_e \dots A_p$ ($p \geqslant 1$), and the incompatibility edges which are incident in any of these vertices constitute a subset of the incompatibility edges which are incident in A in graph G_i, then the chromatic number $K(G_i^*)$ of the new graph G_i^* which is obtained is less than or equal to the chromatic number $K(G_i)$ of G.

Proof - If we color G_i in any way, and then give to $A_1, \dots, A_e, \dots, A_p$ in G_i^* the color of A in G, and to all other vertices in G_i^* the color of the corresponding vertices in G_i, we obtain a coloring of G_i^*. In fact, every vertex in G is given a color, and no two vertices which are joined by an incompatibility edge in G_i^* are colored the same, since the incompatibility edges incident in $A_1 \dots, A_e \dots, A_p$ are subsets of the incompatibility edges incident in A in G_i. Hence $K(G_i) \geqslant K(G_i^*)$.

Q.E.D.

As a consequence of Theorem 2, if we consider a network $N_i \in \mathcal{N}$ having a thickness $\theta(N_i)$, and we introduce a hole point in a path belonging to the network , the new thickness $\theta^*(N_i)$ is less than or equal to $\theta(N_i)$. Thus, also $\theta_{\min}(\mathcal{N}) = \min_{i=1,u} \theta(N_i)$ decreases or remains unchanged.

In the next section, the decomposition algorithm for determining $\theta_{\min}(\mathcal{N})$ and the networks corresponding to it will be illustrated.

IV. THE METHOD OF SOLUTION

In order to determine the network $N_i \in \mathcal{N}$ having the minimum thickness, the incompatibility graph G is firstly determined. Graph G is derived introducing the hole points on the preassigned paths which are not uniplanar, and checking the resulting paths for pairwise compatibility by means of simple geometric considerations.

The junction and the separation relations are then taken into account; the modified graph G' is so obtained. In graph G', the path edges are then extracted, and the redundancy edges are transformed into incompatibility edges, so obtaining a new graph G".

For our example, graph G" is shown in Fig. 4(b).

Let us say a complete subgraph of a given graph to be maximal if it is not subgraph of any other complete subgraph. The next step of the decomposition algorithm is the determination of the set of maximal subgraphs of the complementary graph $\overline{G}''$ of graph G". Recalling Definitions 7 and 8, we conclude that maximal subgraphs of $\overline{G}''$correspond to maximal compatibility classes of paths. Using these classes a solution to our problem will be derived.

Note that while the graphical representation of graph G" is useful in small problems to visualize the relations, it becomes rapidly cumbersome when problems grow large; hence, a tabular representation, which is called incompatibility table, is used. This table is the connectivity matrix of the graph, and, since the incompatibility relation is symmetric, only the entries on one side of the main diagonal need to be considered. In Fig. 5(a) the incompatibility table for the graph G" in Fig. 4(b) is shown.

Some well-known switching theory algorithms exist for determining the maximal subgraphs of $\overline{G}''$, i.e., the set of maximal compatibility classes, starting from the incompatibility table. See, for example, references [5-6] and [10-11] . Using one of these algorithms, in our example the maximal incompatibility classes shown in Fig. 5(b) may be obtained.

Recalling that in graph G" redundant edges are treated as incompatibility edges, it follows that each compatibility class of paths contains no more than one path from each redundant path set. Hence, recalling Defintions 7 and 8, we conclude that maximal compatibility classes of paths constitute maximal sets of paths which can be implemented on the same layer without having unwanted interferences; thus, they constitute also the maximal sets of vertices in graph G' which may be colored the same. Actually, only the maximal compatibility classes need to be found; all other compatibility classes are subsets of these.

The vertices of a subgraph G_i, corresponding to a network $N_i \in \mathcal{N}$, will be covered by at least one subset of maximal compatibility classes. The chromatic number $K(G_i)$ of G_i, i.e., the thickness $\theta(N_i)$ of N_i, will correspond to the cardinality of any smallest subset of maximal compatibility classes whose union contains all the vertices of G_i. Actually, such cardinality constitutes the minimum number of compatibility classes covering all the vertices of G_i; thus, it defines the lower bound on the number of colors in any coloration of G_i.

The problem of determining $\theta_{min}(\mathcal{N})$ is then equivalent to the problem of finding all the smallest subsets of maximal compatibility classes whose union contains at least one path (or one set of elementary paths whose union is a path) from each redundant path set : if among these subsets we select the one(s) having the minimum cardinality, then this cardinality is equal to $\theta_{min}(\mathcal{N})$. These subsets of maximal compatibility classes are said to be minimal covering subsets.

A network $N_i \in \mathcal{N}$, whose paths (or sets of elementary paths whose union is a path) are covered by the union of the maximal classes belonging to a minimal covering subset, has a thickness $\theta(N_i) = \theta_{min}(\mathcal{N})$, and thus it constitutes a solution; such a network is said to be a solution network.

From the above considerations, a switching theory method has been derived for determining the minimal covering subsets. A boolean variable C_h is associated to each maximal compatibility class C_h. Let us consider in graph G' each subset R_k^t of vertices which are joined together by paths edges to form a complete subgraph of order v_k^t. For each vertex $R_k^{t,g}$ belonging to this subset, the boolean sum of all variables associated with compatibility classes having $R_k^{t,g}$ as member is operated. The product of the so obtained sums is then made for $g = 1,\dots v_k^t$, and the resulting boolean expression is associated with R_k^t. For example, consider vertices 5 and 6 in Fig. 4(a), which are joined by a path edge to form a complete graph having order 2. Using the compatibility classes derived in Fig. 5(b), we will build the expression : ((A+B+E) (A+E)), which is equal to (A+E).

It is obvious that, each addend of the resulting expression represents a subset of maximal compatibility classes whose union includes all the members of R_k^t. In our example, these subsets are {A} and {E} , and they both include vertices 5 and 6.

Let us now consider all the subsets R_k^f ($f=1,\dots s_k$) which are members of the same redundant path set. The boolean sum is then made of the previously derived expressions which are associated with each R_k^f, for $f = 1, \dots s_k$. The resulting expression is associated with $\mathcal{R}_k$. In our example, we can consider the two subsets (5,6) and (3,4). We will write :

$$((A+B+E)\cdot(A+E) + (C+D) . (C)) = A+E+C$$

The product of the expressions associated with each R_k is then operated for K=1, ... $\overline{K}$, where $\overline{K}$ is the total number of redundant path sets which are present in graph G'. In our example we obtain : (B+C+E) . (B+C+E) . ((A+D) + (B))$\cdot$((A+B+E).(A+E)+ + (C+D).(C)) = AB+AC+AE+BC+BE+CD+ED.

Each addend of the resulting expression represents a set of maximal compatibility classes whose union contains at least one path (or one set of elementary paths whose union is a path) from each redundant path set. Hence, selecting among this addends the one(s) containing the minimum number of maximal compatibility classes, we obtain the minimal covering subsets. In our examples, all the 7 subsets which have been obtained, (i.e., {A,B} , {A,C} , {A,E} , {B,C} , {B,E} , {C,D} , and {E,D}), are minimal covering subsets. They all have cardinality 2; hence, $\theta_{min}(\mathcal{N}) = 2$. Two layers are then needed to implement the given interconnection network.

A number of solution networks N_i can be derived from the minimal covering subsets. In our example, considering only the minimal covering subset {A,C} , two solution networks can be derived: N_1, which is constituted by paths $(1_a \cup 2_a \cup 3_a \cup 4_a) \equiv (R_1^1)$ (see Fig.1) and $(12_b) \equiv (R_2^1, R_3^1)$, and N_2, which is constituted by paths $(1_a \cup 2_a \cup 5_b \cup 6_b) \equiv (R_1^2)$ and $(12_b^2) \equiv (R_2^1, R_3^1)$; subscrits a and b indicate that the path must be implemented on the first or on the second layer respectively. Other solution networks can be derived from the other 6 minimal covering subsets.

Some criteria can be devised for selecting one among the possible solution networks. For instance, the network which leads to the most efficient or regular distribution of the paths over the different layers could be selected; another criterion could be the minimization of the number of through holes between layers. According to these criteria, between the two solution networks N_1 and N_2 derived in our example, N_1 should be selected, because it requires no through holes, and the paths implemented on each layer have a similar complexity.

V. CONCLUSIONS

The problem discussed here belongs to a class of problems in which, given a finite, but generally large set of elements, those elements which are optimal according to a specified criterion must be determined. In our problem the elements are the possible networks interconnecting n given points, and the optimal elements are the networks having the minimun thickness.

Such problems may always be solved using a complete enumerative procedure; however, this procedure is impractical when the sets are large.

In this paper a direct computer-oriented procedure, which determines the optimal elements without enumeration, has been illustrated. However, since the procedure needs the determination of the maximal compatibility classes, it potentially requires a large amount of computation and memory space when the problem grows large. Usually, the more confining are the constraints involving junction and separation relations between paths, the more likely it is that the number and the cardinality of maximal compatibility classes will be sufficiently limited to allow the problem to be solved.

ACKNOWLEDGMENT

The author is grateful to Dr. Vito Amoia for his critical reading of the paper.

REFERENCES

[1] W.T.TUTTE - "The Thickness of a Graph", Nederl.Akad.Wetensch. Proc.Ser., Vol. 66, pp. 567-577, 1963

[2] A.M.HOBBS - "A Survey of Thickness", in Recent Progress in Combinatorics , Academic Press, New York and London, 1969 (edited by W.T.TUTTE)

[3] M.CHEIN - "Solution of Steiner's Problem in a Particular Case", Automatisme , Vol. XIV, n.10, October 1969 (in French)

[4] J.A.SCARLETT - "Printed Circuit Boards for Microelectronics" Van Nostrand Reinhold Comp., London, 1970

[5] M.C.PAULL and S.H.UNGER - "Minimizing the Number of States in Incompletely Specified Sequential Switching Functions", IRE Trans. on Electronic Computers, Vol. EC-8, pp. 356-367, September 1959

[6] E.J.McCLUSKEY - "Introduction to the Theory of Switching Circuits", Mc Graw Hill, New York, 1965

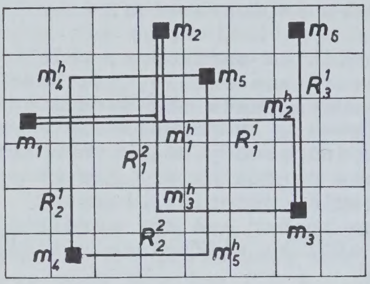
[7] G.A.DIRAC - "Note on the Colouring of Graphs", Mathematische Zeitschrift, Vol. 57, pp. 347-353, 1951

[8] OYSTEIN ORE - "The Four-Color Problem", Academic Press, New York and London, 1967

[9] M.A.GAVRILOV and A.D.ZAKREVSKII (editors) - "LY a PAS: a Programming Language for Logic and Coding Algorithms, Academic Press, New York and London, '69.

[10] J.WEISSMANN - "Boolean Algebra, Map Coloring and Interconnections", Amer. Math. Monthly, Vol.69, pp. 608-613, 1962

[11] A.GRASSELLI - "A Note on the Derivation of Maximal Compatibility Classes", Calcolo, Vol. 3, pp. 165-176, 1966.


m_2
m_6
m_4^h
m_5
R_3^1
m_2^h
m_1
m_1^h
R_1^1
R_1^2
m_3^h
R_2^1
m_3
R_2^2
m_4
m_5^h

Common Sets:

$S_1 = \{m_1, m_2, m_3\}$

$S_2 = \{m_4, m_5\}$

$S_3 = \{m_3, m_6\}$

Redundant Path Sets:

$\mathcal{R}_1 = \{R_1^1, R_1^2\}$

$\mathcal{R}_2 = \{R_2^1, R_2^2\}$

$\mathcal{R}_3 = \{R_3^1\}$

FIG. 1 - SAMPLE BOARD

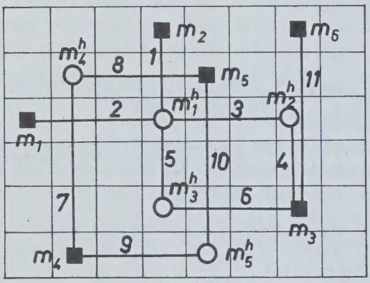

m_2
m_6
m_4^h
8
1
m_5
11
m_2^h
2
m_1^h
3
m_1
5
10
4
m_3^h
6
7
m_3
9
m_4
m_5^h

FIG. 2 - SIMPLE PATHS

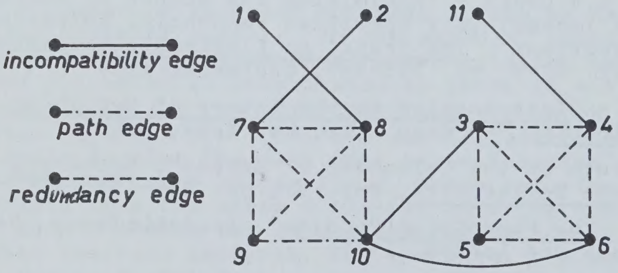

1
2
11
incompatibility edge
path edge
redundancy edge
7
8
3
4
9
10
5
6

FIG. 3 - INCOMPATIBILITY GRAPH G

MEMORY ACCESS CONFLICT RESOLUTION IN MULTIPROCESSING*

J. M. DANIEL
Auburn University
Auburn, Alabama

J. D. IRWIN
Auburn University
Auburn, Alabama

Recent developments in the area of Multiprocessing have led to studies concerned with the efficient use of available memory.

This paper describes a technique for resolving memory access conflicts in a multiprocessor environment. The method, which extends the work of Gordon [1] to the N-processor case, is simple, straight forward and easy to apply. It is assumed that (a) the memory unit is segmented into a number of independent addressable modules and that all processors have random access to each module, and (b) the system operated in an asynchronous mode and generates a start memory cycle signal at the same time that access is granted to a particular processor. In addition, it is shown that the probability of conflict for a given number of memory modules becomes fairly large as the number of central processing units increase. A step-by-step procedure for designing the conflict resolution system is given. The design of this system, which allows for safe race conditions, is based on a decision rule for dynamic assignment of processor priorities. This assignment routine is structured so as to avoid processor lockout; i.e. The prohibiting of one or more processors from ever gaining access to a particular memory location.

An important facet of this conflict resolution system is the ease with which it can be realized with hardware, thus avoiding the need for currently available software solutions for memory protection and access conflicts, which appear to be inefficient due to the large amount of processing time needed for their implementation.

REFERENCES

1. C. R. Gordon, "Dynamic Priority Assignment Circuit," IBM Technical Disclosure Bulletin, vol. 9, pp. 1331-1332, March 1967.

---

*This paper was partially supported by the Army Missile Command, Huntsville, Alabama under Project THEMIS contract DAAH01-68-C-0296.

OPTIMAL CONTROL FOR A CLASS OF NONLINEAR HEREDITARY DYNAMICAL SYSTEMS

N.U. AHMED*
University of Ottawa,
Ottawa, Canada.

INTRODUCTION

In this paper we solve certain problems of existence of optimal controls for a class of Hereditary dynamical systems (HDS) considered more general than the class for which Friedman (2, pp 396-) presented a proof of maximum principle (Pontryagin 5, pp 50,81). Proof of existence of optimal controls for this class is presented under much weaker hypothesis than those used by Friedman (2, pp 412).

The control set is chosen to be a compact and convex subset $U \subset E^r$ containing the origin and the class of bounded measurable functions, defined on compact intervals and taking values in U, is chosen to be the class of admissible controls. The choice of a compact set $U \subset E^r$ in contrast to an arbitrary set (1 pp 281-) for the range of the controls makes it possible to use the w* (weakstar) - topology of L_∞ for the admissible class of controls. By use of this property and few introductory lemmas it is shown that for any fixed initial condition the solution process of the system (HDS) is closed and compact. The attainable sets are shown to be compact subsets of E^n. These results are then used to solve a general optimization problem with applications to standard and nonstandard problems of fuel and time optimal controls.

The problem of optimal choice of an initial position (in the state space) and optimal steering of the system (HDS) into a compact target set is considered after introducing certain generalization of the previous results.

1. BASIC ASSUMPTIONS AND STATEMENT OF THE PROBLEM

We are interested in the problem of optimal control for a class of Hereditary dynamical systems (HDS) described by the following integro-differential equation in E^n.

$$\text{HDS} \qquad \frac{dx(t)}{dt} = f(t,x(t),\, u(t)) + \int_{t_o}^{t} K(t,\tau)\, g(\tau, x(\tau),\, u(\tau))\, d\tau \qquad \text{a.e. on}$$

$I \;\Delta\; [t_o, t_1] \subset [t_o, \infty)$ with $x(t_o) = x_o \in V_o$ where V_o is any closed bounded subset of the Euclidean space E^n.

$x(t) \in E^n$ and $u(t) \in E^r$ for each $t \in I$; $f\colon I \times E^n \times E^r \quad E^n$,

$g\colon\; I \times E^n \times E^r \quad E^n$ and K is a (nxn) matrix valued function on $t_o \le \tau \le t \le t_1$.

Let U be any compact convex subset of E^r containing the origin and $L_\infty(I, E^r)$ the space of essentially bounded Lebesgue measurable functions

---

* This work was supported in part by the National Research Council of Canada under grant No A-7109.

defined on I and taking values in E^r. The following definitions are useful both in the statement of the problem and later.

Definition 1.

For the class of admissible controls we define the set $B(U) \triangleq \{u \in L_\infty(I, E^r) : u(t) \in U$ a.e. on $I\}$. Thus a control u defined on I and taking values in E^r is said to be admissible if $u \in L_\infty(I, E^r)$ and $u(t) \in U$ a.e. (almost everywhere) on I.

Definition 2.

For an arbitrary but fixed $x_o \in V_o$ the set $X \subset C(I, E^n)$ defined by $X \triangleq \{ x \in C(I, E^n) : \dot{x}(t) = f(t, x(t), u(t)) + \int_t^t K(t, \tau)\, g(\tau, x(\tau), u(\tau))\, d\tau$ a.e. on I and $x(t_o) = x_o$, $u \in B(U)\}$ is the set of all motions performed by the system HDS during the time interval I under the action of all possible controls that can be chosen from the set B(U). The set X is called the solution process of the system HDS.

Let $\varphi : B(U) \otimes X \rightarrow R_o = \{r \geq o\}$ be any real valued nonnegative lower semicontinuous function (in an appropriate topology to be defined later) on B(U) x X. The problem is to choose a control $u^* \in B(U)$ so that $\varphi(u^*, x^*) \leq \varphi(u, x)$ for all $(u, x) \in B(U) \times X$ where x^* and x are the responses of the system corresponding to the controls u^* and u respectively. This problem is solved in section 2 after the basic properties of the admissible control set B(U) and the solution process X are developed. In section 3 certain specific applications of these results are made in fuel and time optimal controls. In section 4 the preceding results are generalized and a terminal control problem is solved.

The following assumptions are used throughout the rest of the paper without mention.

A_f : For almost all $t \in I$ f is a continuous function on $E^n \times E^r$ to E^n and for each fixed $z \in E^n$ and $v \in E^r$ it is a measurable function on I to E^n. Further for almost all $t \in I$ and for each $x \in E^n$ $f(t, x, U)$ is a closed convex subset of E^n. In addition, for each fixed $y, z \in E^n$ there exists a $K_1 \in L_1(I)$ and non-negative numbers $\alpha_1, \beta_1 \geq o$ such that

(i) $|f(t, y, u) - f(t, z, u)| \leq |K_1(t)| \; |y - z|$ a.e. (almost everywhere) on I and for all $u \in U$;

(ii) $|f(t, y, u)| \leq |K_1(t)| \; (\alpha_1 + \beta_1 |y|)$ a.e. on I and for all $u \in U$, where the vector spaces E^n and E^r are given the norm $|z| \triangleq \sum_i |z_i|$.

A_g : For almost all $t \in I$ g is a continuous function on $E^n \times E^r$ to E^n and for each fixed $z \in E^n$ and $u \in E^r$ it is a measurable function on I to E^n. In addition for almost all $t \in I$ and for each $x \in E^n$ $g(t, x, u)$ is closed convex subset of E^n. Further for each pair $y, z \in E^n$ there exists a $K_2 \in L_1(I)$ and non-negative numbers $\alpha_2, \beta_2 \geq o$ so that

(i) $|g(t, y, u) - g(t, z, u)| \leq |K_2(t)| \; |y - z|$ a.e. on I for all $u \in U$ and

(ii) $|g(t, y, u)| \leq |K_2(t)| \quad (\alpha_2 + \beta_2 |y|)$ a.e. on I for all $u \in U$.

A_K: The Kernel $K(t, \tau)$ is any essentially bounded Lebesgue measurable ($n \times n$) matrix valued function defined on the triangle $t_o \leq \tau \leq t \leq t_1 < \infty$ and integrable with respect to t on $[\tau, t_1]$ for every $\tau \in I$.

In the following section properties of the solution process X and the attainable set A(t) are developed and a proof of the existence of an optimal control is presented.

2. PROPERTIES OF THE SOLUTION PROCESS X.

Existence and uniqueness of solution of the system (HDS) is presented in the following

Proposition 1.

Under the assumptions A_f, A_g and A_k and for each fixed initial condition $x_o \in V_o \subset E^n$ and a control $u \in B(U)$ the system HDS has one and only one (absolutely continuous) solution $x \in C(I, E^n)$.

Proof : The proof is standard and omitted.

For convenience we will denote the response of the system HDS the control $u \in B(U)$ by the symbol x_u corresponding to an arbitrary but fixed initial vector $x_o \in V_o \subset E^n$.

Lemma 1.

The solution process X of the system HDS is a bounded subset of $C(I, E^n)$.

Proof: (see ref 11).

Definition 3

A subset D of Banach space $C(I, E^n)$ is said to be equicontinuous if for every $\epsilon > o$ there exists a $\delta > o$ such that for all t and $s \in I$ satisfying the inequality $|t-s| < \delta$, $|x(t) - x(s)| < \epsilon$ for all $x \in D$.

Lemma 2.

The solution process X of the system (HDS) is an equicontinuous subset of the space $C(I, E^n)$.

Proof : (see ref 11)

From the above lemma we obtain the following

Corollary 1.

The solution process $X \subset A\,C(I, E^n)$ where $A\,C(I, E^n)$ is the family of absolutely continuous functions defined on I and taking values in E^n.

Definition 4.

A subset D in a Banach space B is called conditionally compact if its closure (in its norm topology) is compact.

We are now prepared to prove the following

Proposition 2.

Under the assumptions A_f, A_g and A_K the solution process X of the system HDS is a conditionally compact subset of $C(I, E^n)$ (in its topology of uniform convergence).

Proof:

Since I is a closed bounded interval in R and by lemma 1 and lemma 2 X is a uniformly bounded equicontinuous subset of $C(I, E^n)$ Arzella-Ascoli theorem (Dunford and Schwartz, th 7, pp 266) holds. From this theorem we obtain that X is conditionally compact.

Definition 5.

By the closure of the solution process X of a dynamical system S it is understood that if $\{x_n\} \in X$ is any sequence of solutions of the system S and x_n x_o (in the relative topology of X induced by the topology of the space in which it is embedded) then x_o is also a solution of the system S.

We will be able to prove in the sequel that X is actually closed in the topology of uniform convergence. We need the following preparation.

Lemma 3.

Let $F: I \times E^r \rightarrow E^n$ so that for each fixed $v \in E^r$, F is a measurable function of t on I taking values in E^n and for almost all $t \in I$, F is a continuous function of v on E^r to E^n. If u is any measurable function on I to E^r then $F(t, u(t))$ is a measurable function on I and if u_n u_o in measure then $F(t, u_n(t)) \rightarrow F(t, u_o(t))$ in measure.

Proof: (see Ref 11).

The above result is a direct generalization of Nemytskii's result (Krasnoselskii, lemma 2.1, pp 20). Let $U \subset E^r$ be an arbitrary compact and convex set containing the origin and I any closed bounded interval on the real line. Let $B(U) \triangle \{u \in L_\infty(I, E^r): u(t) \in U \text{ a.e. on } I\}$ be the class of admissible controls as defined before. Since the dual $L_1^*(I)$ of $L_1(I)$ is $L_\infty(I)$ (ie $L_1^*(I) = L_\infty(I)$) (Dunford and Schwartz, Th 5, pp 289), the L_1 topology of L_∞ defined by the basic neighbourhoods $\{N_\epsilon(x_o)$ $\epsilon \geq o,\ x_o \in L_\infty\}$, where

$$N_\epsilon(x_o) \triangleq \{x \in L_\infty : |z(x-x_o)| = |\int_I z(t)\,[x(t)-x_o(t)]\,dt| < \epsilon, z \in B\}$$

for any finite subset $B \subset L_1(I)$, is the w^* (weak star) topology. Therefore it follows from theorem 2 and corollary 3 (Dunford and Schwartz, pp 424) or theorem 7.1 (La Salle and Hermes, pp 19) that any closed (in the ordinary sense) convex and bounded subset of $L_\infty(I)$ containing the zero element is w^*-compact and w^*-closed. This result is also true for $L_\infty(I, E^r)$ equipped, for instance, with the norm $\|z\| = \sum_{i=1}^{r} \text{ess sup } |z^i(t)|$ for $z \in L_\infty(I, E^r)$. Thus the admissible control set B(U) is w^*-compact and w^*-closed.

In the sequel we need the following

Lemma 4.

If the Lebesgue measure $\mu(I)$ of the set $I \subset R$ is finite then w^* compactness and w^*-closure of the admissible control set $B(U) \subset L_\infty(I, E^r)$ implies that every infinite sequence in B(U) has a subsequence that converges a.e. (almost everywhere) and also in measure on I to its w^* limit.

Proof : (see Ref 11).

Remark:

The converse of the above lemma is also true and can be proved by use of (i) Riesz's theorem that states that every sequence that converges in measure has a subsequence that converges almost everywhere and (ii) Lebesgue dominated convergence theorem. We do not present a detailed proof of this because we do not need it.

We are now prepared to prove the following

Proposition 3.

Under the assumption A_f, A_g and A_k the solution process X of the system HDS is closed (in its uniform topology).

Proof:

Since X is conditionally compact (proposition 2) for every sequence $\{x_n\} \subset X$ there is a subsequence $\{x_{n_k}\} \subset \{x_n\}$ and an element $x^* \in C(I, E^n)$ such that $x_{n_k} \quad x^*$ (uniformly on I). It is required to show that $x^* \in X$. Since $\{x_{n_k}\}_k \subset X$ there exists a sequence of controls $\{u_{n_k}\} \subset B(U)$ so that x_{n_k} is the solution of HDS corresponding to the control u_{n_k} for each $k \in N$ (positive integers). $B(U)$ being w^*-compact and w^*-closed there exists a a subsequence $\{u_{n_{k_s}}\} \subset \{u_{n_k}\}$ and an element $u^o \in B(U)$ so that $u_{n_{k_s}}$ converges to u^o in the w^* topology as $s \quad \infty$. Corresponding to the control $u^o \in B(U)$ let $x^o \in X$ be the unique solution of the system HDS. We must show that $x^* = x^o$. To avoid multiple subscripts it is convenient to define for each $s \in N$ $u_{n_{k_s}} \triangleq v_s$ and $x_{n_{k_s}} \triangleq y_s$. It is clear that $\| x^* - x^o \| \leq \| x^* - y_s \| + \| y_s - x^o \|$ for each $s \in N$ and since $y_s \quad x^*$ (uniformly on I) we have $\| x^* - x^o \| \leq \lim_s \| y_s - x^o \|$.

2.1

Since $\{y_s\} \in X$ corespond to $\{v_s\} \in B(U)$ and $x^o \in X$ corresponds to $u^o \in B(U)$ it follows from equation 2.1 that for every $t \in I$

$$\begin{aligned} | y_s(t) - x^o(t) | \leq & \int_{t_o}^{t} | f(\tau, y_s(\tau), v_s(\tau), v_s(\tau)) - f(\tau, x^o(\tau), v_s(\tau)) | \, d\tau \\ & + \int_{t_o}^{t} | f(\tau, x^o(\tau)) - f(\tau, x^o(\tau), u^o(\tau)) | \, d\tau \\ & + \int_{t_o}^{t} \| R(t,\tau) \| \, | g(\tau, y_s(\tau), v_s(\tau)) - g(\tau, x^o(\tau), v_s(\tau)) | \, d\tau \\ & + \int_{t_o}^{t} \| R(t,\tau) \| \, | g(\tau, x^o(\tau), v_s(\tau)) - g(\tau, x^o(\tau), u^o(\tau)) | \, d\tau \end{aligned} \qquad 2.2$$

Defining $f(t, x^o(t), v(t)) \triangleq F(t, v(t))$, $g(t, x^o(t), v(t)) \triangleq G(t, v(t))$ for each $v \in B(U)$ and for the fixed $x^o \in X$, and using the assumptions $A_f(i)$, $A_g(i)$ and A_k in the inequality 2.2 we obtain for each $t \in I$

$$| y_s(t) - x^o(t) | \leq z_s(t) + \int_{t_o}^{t} h_2(\tau) \, | y_s(\tau) - x^o(\tau) | \, d\tau \qquad 2.2'$$

where $z_s(t) = \int_{t_o}^{t} |F(\tau, v_s(\tau)) - F(\tau, u^o(\tau))| \, d\tau + R \int_{t_o}^{t} |G(\tau, v_s(\tau)) -$

$$- G(\tau, u^o(\tau)) | \, d\tau \, , \qquad 2.3$$

and $h_2(t) = |K_1(t)| + R \, |K_2(t)|$. 2.4

From 2.3 it follows that $z_s(t) \le z_s(t_1)$ for all $t \in I$, therefore

$|y_s(t) - x^o(t)| \le z_s(t_1) + \int_{t_o}^{t} h_2(\tau) \, |y_s(\tau) - x^o(\tau)| \, d\tau$ and

Consequently $\| y_s - x^o \| \le z_s(t_1) \exp \{ \int_{t_o}^{t_1} h_2(\tau) \, d\tau \}$ 2.4'

for all $s \in N$. By 2.4, $h_2 \in L_1(I)$ and therefore the exponential term is finite. We will prove that $\lim_s \| y_s - x^o \| = 0$ by showing that $\lim_s z_s(t_1) = 0$. By the descriptive parts of the hypothesis A_f and A_g it is clear that the functions F and G as defined above satisfy the hypothesis of lemma 3. Further by lemma 4, w^*-convergence of v_s to u^o implies its convergence in measure to the same limit and consequently by lemma 3.

(a) $F(t, v_s(t)) \longrightarrow F(t, u^o(t))$

in measure on I.

and $G(t, v_s(t)) \longrightarrow F(t, u^o(t))$

From the definitions of F and G and the conditions A_f(ii) and A_g(ii) we have for any measurable set $E \subset I$

$\int_E |F(t, v_s(t))| \, dt \le \int_E |K_1(t)| \quad (\alpha_1 + \beta_1 \, |x^o(t)|) \, dt$

2.5

and $\int_E |G(t, v_s(t))| \, dt \le \int_E |K_2(t)| \quad (\alpha_2 + \beta_2 \, |x^o(t)|) \, dt$.

Since X is a bounded subset of $C(I, E^n)$ and $x^o \in X$,

$\int_E |F(t, v_s(t))| \, dt \le (\alpha_1 + \beta_1 b) \int_E |K_1(t)| \, dt$ and $\int |G(t, v_s(t))| \, dt \le$

$(\alpha_2 + \beta_2 b) \int_E |K_2(t)| \, dt$ for some $0 < b < \infty$.

(b) $\lim_{\mu(E) \to o} \int_E |F(t, v_s(t))| \, dt = 0$

for all $s \in N$.

and $\lim_{\mu(E) \to o} \int_E |G(t, v_s(t))| \, dt = 0$

Therefore the sequence $\{F(t, v_s(t))\}$ and $\{G(t, v_s(t))\}$ satisfy all the necessary conditions (a) and (b) of theorem C (Halmos, th C, pp 108) and from this theorem we conclude that

$\int_I |F(t, v_s(t)) - F(t, u^o(t))| \, dt \to 0$

and $\int_I |G(t, v_s(t)) - G(t, u^o(t))| \, dt \to 0$.

Thus $z_s(t_1) \quad o$ (in fact $z_s(t) \quad o$ for all $t \in I$) and from (2.4) $\| y_s - x^o \| \rightarrow o$ and consequently from (2.1) we have $x_* = x_o$.
This proves the closure of X.

Remark: Propositions 2 and 3 imply that the set X is compact (and closed) in the topology of uniform convergence.

Now we are prepared to solve the optimal control problem stated before.

Proposition 4:

Let $\varphi : B(U) \otimes X \quad R_o \triangleq \{ r \geq o \}$. Suppose for each fixed $u \in B(U)$, φ is lower semicontinuous on X with respect to the topology of uniform convergence and for each fixed $x \in X$ it is lower semicontinuous on B(U) in its w^*-topology and that it is jointly lower semicontinuous in the product topology. Then there exists a control $u^* \in B(U)$ and a corresponding trajectory $x^* \in X$ so that $\varphi(u^*, x^*) \leq \varphi(u,x)$ for all $(u,x) \in B(U) \otimes X$ with $x = x_u$.

Proof:

Define $D = \{ (u, x_u) : u \in B(U) \} \subset B(U) \otimes X$.

Let $\{ u_n \} \in B(U)$ be a sequence of controls and $\{ x_n \} \in X$ the corresponding sequence of trajectories. If $u_n \quad u_o$ in the w^*-topology then, B(U) being w^*-closed, $u_o \in B(U)$ and similarly if $x_n \quad x_o$ (uni) then, by proposition 3, $x_o \in X$ and $x_o = x_{u^o}$.

Therefore, D is a closed subset of $B(U) \otimes X$. But, since the w^*-topology on B(U) and the "uniform" topology on X are both Hausdorf topologies, $B(U) \otimes X$ is a compact Hausdorf space in the product topology. D being a closed subset of a compact Hausdorf space is also compact. Thus the function φ, being lower semicontinuous in the product topology, attains a lower bound on D. The proof is similar to that for theorem 9.2 (Vainberg, th 9.2 pp 78). Thus there is a pair $(u^*, x^*) \in D$ so that $\varphi(u^*, x^*) \leq \varphi(u,x)$ for all $(u,x) \in D$ where $x = x_u$ and $x^* = x_{u^o}$. This completes the proof of the proposition.

Definition 6.

For each $t \in I$ the set $A(t) \triangleq \{ x(t) \in E^n : x \in X \}$ is called the attainable set. In the notation of Varayia and Eggert (1, pp 280-299, 1968) $A(t) = T(x_o, t_o t)$ - the set of states that can be attained by the system at time t starting from the state x_o at time t_o.

Corollary 2

The attainable sets $\{ A(t), t \in I \}$ are compact subsets of E^n.

Proof:

By lemma 1, for reach $t \in I$, A(t) is bounded and by proposition 3 it is closed. Therefore A(t) is compact for each $t \in I$.

3. APPLICATIONS

(A) FUEL OPTIMAL CONTROL

Consider the system HDS and the cost functional $\varphi(u,x) = \int_{t_o}^{t_1} |Mx| dt + \int_{t_o}^{t_1} |Nu| \, dt + \varphi_o(x(t_1))$ where φ_o is any lower semicontinuous function on E^n to R_o, M is any (mxn) matrix with elements $m_{ij}(t) \in L_1(I)$ and N is any (sxr) matrix with elements $n_{ij}(t) \in L_1(I)$. The problem is to choose a control $u \in B(U)$ that minimizes the function φ. By proposition 4 it is enough to show that the cost functional φ is lower semicontinuous on $B \otimes X$. For definite see ref 11.

Remark: The above result is also true for the more frequently used cost functional $\varphi(u,x) = \int_{t_o}^{t_1} (x'Mx)dt + \int_{t_o}^{t_1} (u'Nu)\,dt + (x'(t_1)Px(t_1))$where M and N are required to be merely positive semi-definite for almost all $t \in I$ with Lebesgue integrable entries and P is a positive semi-definite constant matrix. This follows from the facts that (i) $(x'Mx) \leq \|x\| \, |Mx| \leq b \, |Mx|$ a.e. (ii) $(u'Nu) \leq \|u\| \, |Nu| \leq c \, |Nu|$ a.e and $(x'(t_1)Px(t_1)) \leq |x(t_1)| \, |Px(t_1)| \leq b \, |Px(t_1)|$ for all $x \in X$ and $u \in B(U)$ where X and U determine b and c respectively.

B TIME OPTIMAL CONTROL

Let H be any closed subset of E^n. The problem is to choose a control $u^* \in B(U)$ that transfers the system from the state $x_o \notin H$ to the set H in minimum time. Define $X_o \underline{\triangle} \{ x \in X: x(t) \in H$ for some $t \in I \}$. Clearly $X_o \subset X$ is the set of all trajectories of the system HDS that intersect H. The set X_o may be empty for a fixed compact interval $I \subset [t_o, \infty)$ then the problem is trivial and either it has no solution or $\tau_H = \infty$. Therefore we may assume that X_o is non-empty and we obtain the following

Lemma 5.

If the set $X_o \subset X$ is non-empty then it is closed and since X is bounded X_o is also bounded and therefore compact.

Proof: (see ref 11).

The above result is also true for an arbitrary continuous set valued function H(t) defined for $\infty > t \geq$ to and taking values in the Hausdorf metric space of compact sets in E^n.

Proposition 5.

If the set X_o is non-empty then there exists a control $u^* \in B(U)$ that transfers the system from the initial state x_o ($x_o \notin H$) to the set $H \subset E^n$ in minimum time.

Proof: (see ref 11).

4. CERTAIN GENERALIZATIONS

Let V_o be an arbitrary non-empty compact subset of the real Euclidean space E^n and B(U) the class of admissible controls as defined before (Definition 1.) and let $x_{u,x_o} \in C(I, E^n)$ be the solution of the system HDS corresponding to the initial condition $x_o \in V_o$ and the control $u \in B(U)$. Define $Y = \{ x_{u,w} \in C(I, E^n): w \in V_o, u \in B(U) \}$. Clearly $Y = \bigcup_{w \in V_o} X_w$ where X_w is the solution process of the system HDS corresponding to the initial condition $w \in V_o$ (Definition 2).

The following proposition is an extension of proposition 3.

Proposition 6.

The set $Y \subset C(I, E^n)$ is conditionally compact and closed (and therefore compact).

Proof. (see ref 11)

From this proposition one can easily obtain the following

Corollary 3.

If the set $V_o \subset E^n$ is compact then the set $V_1 \triangleq \{x(t_1) \in E^n : x \in Y\}$ called the set of attainability (of the system HDS at time t_1) is also compact.

There is considerable practical interest in the following optimal control problem. Consider the system

$$\text{HDS} \qquad \dot{x}(t) = f(t,x(t),u(t)) + \int_{t_o}^{t} K(t,\tau)g(\tau,x(\tau),u(\tau))\,d\tau \quad \text{a.e. on } I$$

$$I = [t_o, t_1], \; -\infty < t_o < t_1 < \infty$$

along with (i) a given set $V_o \subset E^n$ as the set of admissible intial conditions (ii) a set $D_1 \subset E^n$ as the set of desirable final states (iii) a set $B(U) \subset L_\infty(I, E^r)$ (as defined before) as the set of admissible controls and (iv) a cost function $\psi(x(t_1))$ defined on D_1. The problem is to choose an initial state $x_o^* \in V_o$ and a control $u^* \in B(U)$ so that $x^*(t_1) \in D_1$ where $x^* = x_{u^*, x_o^*}$ and $\psi(x^*(t_1))$ is a minimum. This problem is solved in the following

Corollary 4.

If the sets V_o and $D_1 \subset E^n$ are compact, the set $D_1 \cap V_1$ is non-empty and the function ψ defined on D_1 is lower semicontinuous then there exists an optimal pair $(x_o^*, u^*) \in V_o \otimes B(U)$ that yields a minimum to ψ on D_1.

Proof.

If the set V_o is compact then by Corollary 3, V_1 is also compact therefore $D_1 \cap V_1$ is a compact subset of E^n whenever it is nonempty and D_1 is compact. ψ being a lower semicontinuous function on D_1 it attains a minimum on the compact set $D_1 \cap V_1$. Define $Y_1 \triangleq \{x \in Y : x(t_1) \in D_1\}$; Clearly $D_1 \cap V_1$ is non-empty if and only if Y_1 is non-empty Suppose $D_1 \cap V_1$ is non-empty and at $z \in D_1 \cap V_1$ ψ attains its minimum then there exists an $x^* \in Y_1$ such that $x^*(t_1) = z$ and consequently there exists a pair $(x_o^*, u^*) \in V_o \otimes B(U)$ so that $x^* = x_{u^*, x_o^*}$. This completes the proof.

Remarks:

(i) It appears to the author that, with slight modification, the results of this paper may be shown to be true for the system

$$\dot{x}(t) = f(t,x(t),u(t)) + \int_{t_0-T}^{t} K(t,\tau)\, g(\tau,x(\tau),u(\tau))d\tau \text{ a.e on } I \triangle [t_0,t_1],$$

with past conditions: (a) $x(t) = x'(t)$, for $t \in [t_0 - T, t_0]$ and (b) $u(t) = u'(t)$ for $t \in [t_0 - T, t_0]$. The past data x' and the past control u' may be either fixed or chosen optimally from an appropriate class.

(ii) Introducing some minor modifications it is possible to prove a maximum principle for our class of systems (HDS) along the same line as Friedman (2, pp 396-416) and Pontryagin et al (5, th 3, 8; pp 50, 81) provided the Kernel K of the system is of the form $K = K(t-\tau, \tau)$. For a general Kernel $K = K(t,\tau)$ a proof of the maximum principle will require further modifications of Pontryagin's approach utilized by Friedman.

(iii) It is interesting to mention that if the Kernel K(of the system HDS) $\equiv 0$ (that is if the hereditary part is absent) and the control set $U \subset E^r$ is compact then it follows from proposition 3 of this paper and theorem 7.3 (1, pp 298) that the dynamical system (for $K \equiv 0$), the corresponding "relaxed dynamical system" (in the sense of Warga) (1, pp 287) and the "derived generalized dynamical system" (1, pp 285) are all equivalent in the sense that $R = X = M$. In Eggert and Varaiya's notation $X = F$.

REFERENCES

(1) D. Eggert and P. Varaiya, "Representation of a Differential System", Journal of Differential Equations 4, pp 280-299, 1968.

(2) A. Friedman, "Optimal control for Hereditary processes" Rational Mechanics and Analysis, Vol. 15, pp 396-416, 1964.

(3) H. Hermes and J.P. La Salle, "Functional Analysis and Time Optimal Control" Ac. Press, pp 19, 1969.

(4) M.N. Oguztoreli, "Time lag control systems", Ac. Press, pp 181, 1966.

(5) L.S. Pontryagin, V.G. Boltyanski, R.V. Gamkrelidze and E.F. Mishchenko; "The Mathematical theory of optimal processes" (Eng Trans) J. Wiley; pp 50, 80; 1962.

(6) N. Dunford and J.T. Schwartz, "Linear operators, part 1", pp 266, pp 424, 1964.

(7) P.R. Halmos, "Measure theory" Van Nostrand, pp 88, pp 92, pp 108, 1964.

(8) M.M. Vainberg, "Variational Methods for the study of Nonlinear Operators" Holden Day, pp 78, 1964.

(9) M.A. Krasnoselskii, "Topological methods in the theory of nonlinear integral equations" Eng Trans. Pergamon Press, pp 20, 1964.

(10) N.U. Ahmed, "Optimal control of a class of nonlinear systems on Hilbert space" IEEE Trans AC-14 No 6, pp 711-714, 1969.

(11) N.U. Ahmed, "Optimal Control for Hereditary Dynamical Systems" TR 70-8, University of Ottawa, Canada.

PULSE MODULATED OPTIMAL CONTROLLERS FOR TIME LAG SYSTEMS

J. WANINSKI and A. G. VACROUX
Department of Electrical Engineering
Illinois Institute of Technology
Chicago, Illinois 60616

ABSTRACT

Necessary conditions are derived for the optimal pulse modulated control of multivariable systems with delays depending on time and state. The conditions are obtained in the most general form for aperiodically modulated control signals. They are applied to the particular cases of PA, PW, PA-PW and PPM controllers.

INTRODUCTION

When a digital computer is used as a control element it is often better to consider it as an aperiodically modulated controller rather than a continuous one, especially in the case of a time shared operation. Likewise, systems with pulse modulated communications channels have piecewise constant controllers. This paper investigated such systems when, in addition, delays are allowed.

Jordan and Polak [1] developed a method analagous to Pontryagin's[2] to treat the problem of piecewise constant optimal control of systems described by differential equations. Extension to retarded systems were considered by Becker and Lopresti[3] and by Banas and Vacroux[4]. Kirk[5] and Nardizzi and Bekey[6] have used the calculus of variations to treat optimal control systems with various forms of modulated controls.

This paper, using a method paralleling Pontryagin's, considers the pulse modulated optimal control of systems described by non-linear differential-difference equations. The system state vector is considered to have a delay which may be a function of both time and state. A theorem is derived which gives the necessary conditions for optimality when the control signals belong to the general class of aperiodically modulated functions. The control amplitudes may be restricted and the switching times may be fixed or variable. The application of the theorem is considered for various pulse modulated controllers including PA, PW, PA-PW and PPM control systems. It is finally shown that a gradient technique can be employed to synthesize the optimal controls. An example problem is given which has been solved using the digital computer and the results are discussed.

FORMULATION OF THE PROBLEM

System In general, the processes under consideration can be described by a non-linear differential-difference equation of the form,

$$\tilde{\dot{x}}_1(t) = \tilde{f}_1[\tilde{x}_1(t),\ \tilde{x}_1(t-\theta(\tilde{x}_1(t),\ t)),\ \tilde{u}(t),\ t],\quad t\varepsilon[t_0,\ t_K] \qquad (1)$$

with the given initial function

$$\tilde{x}_1(t) = \tilde{\phi}(t),\quad t\varepsilon[t_0-\theta(\tilde{x}_1(t_0),\ t_0),\ t_0]$$

where $\tilde{x}_1 \varepsilon E^n$ is an nx1 state vector, $\tilde{u}\varepsilon V$ (a convex subset of E^r) is an rx1 control vector and $\tilde{f}$ is a vector function with components of class C^1 over $E^n \times E^n \times V \times R$. The final state $\tilde{x}_1(t_K)$, with t_K fixed, lies in the set $S = E^n$ (free endpoint problem).

The delay $\theta(\tilde{x}_1(t),t)$ is a given scalar function, continuous and continuously differentiable with respect to its arguments, satisfying $t \geqslant t - \theta(\tilde{x}_1(t), t) \geqslant t_0 - \theta(\tilde{x}_1(t_0), t_0)$ and $\dot{\theta}(\tilde{x}_1(t), t) < 1$ for all t. The first condition ensures that only known past values of state affect the future state. The second is imposed by mathematical considerations.

Admissible Controls Let M be the set of vectors, $\tilde{\mu} = (\tilde{m}_0, \tilde{m}_1, \cdots, \tilde{m}_{K-1}) \ni \tilde{m}_k \varepsilon V$, representing the sequence of control amplitudes. Let T be the set of vectors, $\tilde{\tau} = (t_1, \cdots, t_{K-1}) \ni t_0 < t_1 \leqslant \cdots \leqslant t_{K-1} < t_K$, representing the sequence of switching times. Then $\tilde{u}(t)$ is defined by

$$\tilde{u}(t; \tilde{\mu}, \tilde{\tau}) = \tilde{m}_k, \; t \varepsilon [t_k, t_{k+1}], \; k = 0, 1, \cdots, K-1 \tag{2}$$

$$\tilde{u}(t_K; \tilde{\mu}, \tilde{\tau}) = \tilde{m}_{K-1}$$

Index of Performance To each admissible control there corresponds a cost

$$J[\tilde{u}] = \int_{t_0}^{t_K} f_0[\tilde{x}_1(t), \tilde{u}(t)]dt \tag{3}$$

where the scalar function f_0 is of class C^1 on $E^n \times \mathbf{V}$. Defining an auxiliary variable x_0 which satisfies

$$\dot{x}_0(t) = f_0[\tilde{x}_1(t), \tilde{u}(t)], \; t \varepsilon [t_0, t_K] \tag{4}$$

$$x_0(t_0) = 0$$

it is clear that $J[\tilde{u}] = x_0(t_K; \tilde{u})$.

Statement of the Control Problem For the system (1) it is required to find an admissible control

$$\tilde{u}^*(t) = \tilde{u}(t; \tilde{\tau}^*, \tilde{\mu}^*) \ni x_0(t_K; \tilde{u}^*) \leqslant x_0(t_K; \tilde{u}) \text{ for all } \tilde{u} \varepsilon V.$$

NECESSARY CONDITIONS FOR AN OPTIMAL CONTROL

For notational convenience, system (1) is augmented with the cost variable x_0 to obtain the (n+1) dimensional system

$$\dot{\tilde{x}}(t) = \tilde{f}[\tilde{x}(t), \tilde{x}_\theta(t), \tilde{u}(t), t], \; t \varepsilon [t_0, t_K] \tag{5}$$

$$\tilde{x}(t) = \text{col}(0, \tilde{\phi}(t)), \; t \varepsilon [t_0 - \theta(\tilde{x}(t_0), t_0), t_0]$$

where $\tilde{x} = \text{col}(x_0, \tilde{x}_1)$ and $\tilde{x}_\theta(t) = \tilde{x}(\sigma) = \tilde{x}(t - \theta(\tilde{x}(t), t))$.

The Hamiltonian, H, is defined as

$$H[\tilde{x}, \tilde{x}_\theta, \tilde{p}, \tilde{u}, t] = \tilde{p}^T(t)\tilde{f}[\tilde{x}, \tilde{x}_\theta, \tilde{u}, t] \tag{6}$$

Costate System The (n+1) x 1 costate vector, $\tilde{p}$, is defined as the solution of the differential-difference equation,

$$\dot{\tilde{p}}(t) = - \frac{\partial H}{\partial \tilde{x}}(t) - \frac{\partial H}{\partial \tilde{x}_\theta}(t)\dot{\tilde{x}}^T(\sigma) \frac{\partial \theta}{\partial \tilde{x}}(t) \tag{7}$$

$$- \frac{\partial H}{\partial \tilde{x}_\theta}(\rho) \frac{1}{1 - \dot{\theta}(\rho)}, \; t \varepsilon [t_0, t_K - \theta(\tilde{x}(t_K), t_K)]$$

$$\dot{\tilde{p}}(t) = - \frac{\partial H}{\partial \tilde{x}}(t) - \frac{\partial H}{\partial \tilde{x}_\theta}(t)\dot{\tilde{x}}^T(\sigma) \frac{\partial \theta}{\partial \tilde{x}}(t), \; t \varepsilon [t_K - \theta(\tilde{x}(t_K), t_K), t_K]$$

where the argument $\rho(t)$ is the solution of
$t = \rho - \theta(\tilde{x}(\rho), \rho)$.

Theorem: General Necessary Conditions If, for system (5), $\tilde{u}^*(t) = \tilde{u}(t;\ \tilde{\mu}^*,\ \tilde{\tau}^*)$ is an optimal control with corresponding optimal trajectory $\tilde{x}^*(t) = \tilde{x}(t;\ 0,\ t_0,\ \tilde{u}^*)$, $t\varepsilon[t_0,\ t_K]$, then there exists a non-zero function $\tilde{p}^*(t)$ satisfying the costate equation (7) along the optimal control and trajectory such that

(I) For each $k = 1, \cdots, K$

$$\int_{t_{k-1}}^{t_k} \frac{\partial H}{\partial \tilde{u}} [\tilde{x}^*(s),\ \tilde{x}^*_\theta(s),\ \tilde{p}^*(s),\ \tilde{m}^*_{k-1},\ s]ds\ \delta\tilde{m}_{k-1} \leqslant 0$$

This is a statement that $\int_{t_{k-1}}^{t_k} H[\cdot,\ \tilde{m}^*_{k-1},\ s]ds$ has a local maximum or is a stationary point with respect to $\tilde{m}_{k-1}$.

(II) For each $k = 1, \cdots, K - 1$

$$\left\{H[\tilde{x}^*(t^*_k),\tilde{x}^*_\theta(t^*_k),\tilde{p}^*(t^*_k),\tilde{m}^*_{k-1},t^*_k] - H[\tilde{x}^*(t^*_k),\tilde{x}^*_\theta(t^*_k),\tilde{p}^*(t^*_k),\tilde{m}^*_k,t^*_k]\right\} \delta t_k \leqslant 0$$

If t_k is unrestrained this must hold for both positive and negative δt_k. In this case it is a statement that $H^*[\cdot,\ t^*_k]$ is continuous across the switching instants.

(III) $p_0(t_K) \leqslant 0$

These three conditions are basic and hold regardless of the terminal constraint, the form of which determines the remaining transversality conditions. In our case $S = E^n$ (free endpoint problem), for which the remaining condition is

(IV) $p_i(t_K) = 0,\ i = 1, \cdots, n$

The derivation of the transversality conditions for other forms of the terminal constraint set S is similar to that of Jordan and Polak[1].

PROOF OF THE THEOREM

Perturbations in the Optimal Control Assume the existence of an optimal control $\tilde{u}^*(t) = \tilde{u}(t;\ \tilde{\mu}^*,\ \tilde{\tau}^*)$ and the corresponding optimal trajectory $\tilde{x}^*(t) = \tilde{x}(t,\ 0;\ t_0,\ \tilde{u}^*)$. By the properties of the control set there exists a $\delta\tilde{u} = (\delta\tilde{m}_0, \cdots, \delta\tilde{m}_{K-1}) \neq 0$ and $\delta\tilde{\tau} = (\delta t_1, \cdots, \delta t_{K-1}) \neq 0$ such that, for an $\varepsilon > 0$ sufficiently small, $(\tilde{u}^* + \varepsilon\delta\tilde{\mu})\varepsilon M$ and $(\tilde{\tau}^* + \varepsilon\delta\tilde{\tau})\varepsilon T$. Let $\tilde{u}(t;\varepsilon)$ denote the perturbed control $\tilde{u}(t;\ \tilde{\mu}^* + \varepsilon\delta\tilde{\mu},\ \tilde{\tau}^* + \varepsilon\delta\tilde{\tau})$ and let $\tilde{x}(t;\ \varepsilon)$ denote the perturbed trajectory $\tilde{x}(t;\ 0,\ t_0,\ \tilde{u}(t;\ \varepsilon))$. The variation in the optimal trajectory is given by $\delta\tilde{x}(t;\ \varepsilon) = \tilde{x}(t;\ \varepsilon) - \tilde{x}^*(t)$.

Variational Relations First consider the particular perturbation of the optimal control such that $\delta\tilde{\tau} = \tilde{0}$, $\delta\tilde{m}_i = \tilde{0}$, $i = 0, \cdots, k-2, k, \cdots, K-1$. Then

$$\left[\frac{\partial\tilde{u}}{\partial\varepsilon}(t;\ \varepsilon)\right]_{\varepsilon = 0} = \begin{cases} \delta\tilde{m}_{k-1}, & t\varepsilon[t_{k-1},\ t_k) \\ \tilde{0}, & \text{otherwise} \end{cases} \tag{8}$$

and if $\delta\tilde{x}(t;\ \varepsilon)$ is expanded in a Taylor series about $\varepsilon = 0$, the first term $\tilde{y}(t) = \left[\frac{\partial\tilde{x}}{\partial\varepsilon}(t;\ \varepsilon)\right]_{\varepsilon = 0}$ satisfies the linear differential-difference equation

$$\dot{\tilde{y}}(t) = \left[\frac{\partial\tilde{f}}{\partial\tilde{x}}(t) - \frac{\partial\tilde{f}}{\partial\tilde{x}_\theta}(t)\,\frac{\partial\tilde{x}(\sigma)}{\partial\sigma}\left\{\frac{\partial\theta}{\partial\tilde{x}}(t)\right\}^T\right]_*\tilde{y}(t) + \left[\frac{\partial\tilde{f}}{\partial\tilde{x}_\theta}(t)\right]_*\tilde{y}_\theta(t) + \left[\frac{\partial\tilde{f}}{\partial\tilde{u}}(t)\right]_*\left[\frac{\partial\tilde{u}}{\partial\varepsilon}(t;\,\varepsilon)\right]_{\varepsilon = 0,} \qquad t\epsilon[t_0,\, t_K] \tag{9}$$

$$\tilde{y}(t) = \tilde{0}, \quad t\epsilon[t_0 - \theta(\tilde{x}(t_0),\, t_0),\, t_0]$$

where $[\ \ ]_*$ indicates evaluation for the optimal control and trajectory.

An equation of this form has the solution,

$$\tilde{y}(t) = \int_{t_0}^{t}\tilde{Y}(s,\, t)\left[\frac{\partial\tilde{f}}{\partial\tilde{u}}(s)\right]_*\left[\frac{\partial\tilde{u}}{\partial\varepsilon}(s;\,\varepsilon)\right]_{\varepsilon = 0} ds, \quad t\epsilon[t_0,\, t_K] \tag{10}$$

where the $(n+1) \times (n+1)$ matrix $\tilde{Y}(s,\, t)$ satisfies the adjoint equation

$$\frac{\partial}{\partial s}\tilde{Y}(s,\, t) = -\tilde{Y}(s,\, t)\left[\frac{\partial\tilde{f}}{\partial\tilde{x}}(s) - \frac{\partial\tilde{f}}{\partial\tilde{x}_\theta}(s)\dot{\tilde{x}}(\sigma)\left\{\frac{\partial\theta}{\partial\tilde{x}}(s)\right\}^T\right]_* - \tilde{Y}(\rho,\, t)\left[\frac{\partial\tilde{f}}{\partial\tilde{x}_\theta}(\rho)\,\frac{1}{1 - \dot{\theta}(\rho)}\right]_*, \quad s\epsilon[t_0,\, t - \theta(\tilde{x}(t),\, t)] \tag{11}$$

$$\frac{\partial}{\partial s}\tilde{Y}(s,\, t) = -\tilde{Y}(s,\, t)\left[\frac{\partial\tilde{f}}{\partial\tilde{x}}(s) - \frac{\partial\tilde{f}}{\partial\tilde{x}_\theta}(s)\,\dot{\tilde{x}}(\sigma)\left\{\frac{\partial\theta}{\partial\tilde{x}}(s)\right\}^T\right]_* \quad s\epsilon[t - \theta(\tilde{x}(t),\, t),\, t]$$

with the initial condition $\tilde{Y}(t,\, t) = \tilde{I}$.

Substituting (8) into (10),

$$\tilde{y}(t_K) = \int_{t_{k-1}}^{t_k}\tilde{Y}(s,\, t_K)\left[\frac{\partial\tilde{f}}{\partial\tilde{u}}(s)\right]_*\delta\tilde{m}_{k-1}ds \tag{12}$$

and $\delta\tilde{x}(t_K;\,\varepsilon) = \varepsilon\tilde{y}(t_K) + \tilde{0}(\varepsilon^2)$ where $\lim\limits_{\varepsilon\to 0}\frac{\tilde{0}(\varepsilon^2)}{\varepsilon} = \tilde{0}$.

Now consider a perturbation in the optimal control such that $\delta\tilde{u} = \tilde{0}$ and $\delta t_i = 0$, $i = 1, k-1, k+1, \cdots, K-1$. Then it can be shown that

$$\tilde{y}(t_K) = \tilde{Y}(t_k^*, t_K)\left\{\tilde{f}[\tilde{x}^*(t_k^*), \tilde{x}_\theta^*(t_k^*), \tilde{m}_{k-1}^*, t_k^*] - \tilde{f}[\tilde{x}^*(t_k^*), \tilde{x}_\theta^*(t_k^*), \tilde{m}_k^*, t_k^*]\right\}\delta t_k \tag{13}$$

The Terminal Cone Let C be the set of all vectors $\tilde{x}^*(t_K) + \varepsilon\tilde{y}(t_K)$ where $\tilde{y}(t_K)$ is given by (12) or (13). Since $\tilde{y}(t_K)$ is linear in the control perturbation, C is a convex cone with vertex $\tilde{x}^*(t_K)$. Since $\tilde{u}^*(t)$ is optimal, all points in C formed by the use of perturbed controls have a cost component $x_0(t_K) > x_0^*(t_K)$. Thus C must be separated from the half-space $G = \{\tilde{x} \mid x_0 < x_0^*\}$ by the plane $F = \{\tilde{x} \mid x_0 = x_0^*\}$. An $(n+1) \times 1$ vector in G which is transverse to F is given by $\tilde{a} = \mathrm{col}(a_0, 0, \cdots, 0)$ $a_0 \leqslant 0$. Then $\tilde{a}^T\tilde{y}(t_K) \leqslant 0$. Let $\tilde{p}^*(t_K) = \tilde{a}$, establishing the transversality conditions, III and IV of the theorem. Then,

$$\tilde{p}^{*T}(t_K)\tilde{y}(t_K) \leqslant 0. \tag{14}$$

The Basic Necessary Conditions Combining (12) and (14) gives,

$$\int_{t_{k-1}}^{t_k}\tilde{p}^{*T}(t_K)\tilde{Y}(s,\, t_K)\left[\frac{\partial\tilde{f}}{\partial\tilde{u}}(s)\right]_*\delta\tilde{m}_{k-1}\,ds \leqslant 0 \tag{15}$$

Define

$$\tilde{p}(t) = \tilde{Y}(t,\, t_K)\tilde{p}^*(t_K) \tag{16}$$

and recall the definition of the Hamiltonian (6). Then $\tilde{p}(t)$ satisfies the costate system (7) along the optimal control and trajectory with the boundary condition $\tilde{p}(t_K) = \tilde{p}^*(t_K)$. Now inequality (15) becomes

$$\int_{t_{k-1}}^{t_k} \frac{\partial H}{\partial \tilde{u}} [\tilde{x}^*(s), \tilde{x}_\theta^*(s), \tilde{p}^*(s), \tilde{m}_{k-1}^*, s] ds \; \delta\tilde{m}_{k-1} \leqslant 0 \qquad (17)$$

and since this holds for all $k = 1, \cdots, K$, condition (I) of the theorem is established.

Combining (13) and (14) leads to the inequality

$$\left\{H[\tilde{x}^*(t_k^*), \tilde{x}_\theta^*(t_k^*), \tilde{p}^*(t_k^*), \tilde{m}_{k-1}^*, t_k^*] - H[\tilde{x}^*(t_k^*), \tilde{x}_\theta^*(t_k^*), \tilde{p}^*(t_k^*), \tilde{m}_k^*, t_k^*]\right\} \delta t_k \leqslant 0 \qquad (18)$$

for all $k = 1, \cdots, K-1$. Thus condition (II) of the theorem is established.

PULSE MODULATED CONTROLLERS

The theorem was formulated for a general aperiodically modulated control signal. By introducing additional constraints, the necessary conditions for optimizing systems with various forms of pulse modulated controllers can be obtained.

PAM Control For pulse amplitude modulated control fixed intervals are specified during which the signal amplitude can assume any value in the control set V. These intervals usually span equal lengths of time, although it is not necessary that they do. Since all the switching times are specified, condition (II) of the theorem is deleted. Condition (I) is used to determine the control signal amplitudes.

PWM Control With pulse width modulated control (also called pulse duration modulation) there is one variable switching point within each fixed interval. In the first portion of each interval $[t_{2k-2}, t_{2k-1})$, $k = 1, 2, \cdots, \frac{K}{2}$, the pulse amplitude is restricted to to $+\tilde{A}$, $\tilde{0}$ or $-\tilde{A}$ and in the second portion $[t_{2k-1}, t_{2k})$ the amplitude is zero. The variable switching points, designated as the odd times t_{2k-1}, are determined from condition (II). Since the control amplitudes can assume only certain values, condition (I) is not applicable. A search technique must be used to determine the sequence of amplitudes which is optimal.

PA-PWM Control This is a form of combined modulation differing from PWM in that during the first portion of each interval $[t_{2k-2}, t_{2k-1})$ the pulse amplitude may assume any value in V. Condition (I) is used to determine these amplitudes. The switching points, t_{2k-1}, are determined from condition (II).

PPM Control With a pulse position modulated control a pulse of fixed width with an amplitude of $+\tilde{A}$, $\tilde{0}$ or $-\tilde{A}$ is allowed to move within a fixed interval. Since the switching times at the leading and trailing edges of the pulse are not independent, a modified necessary condition for determining the switching points is obtained. Let t_{2k-1} be the switching time of the trailing edge of the pulse and let w be the width of the pulse. It can be shown that equation (13) of the theorem proof becomes

$$\tilde{y}(t_K) = \tilde{Y}(t^*_{2k-1}, t_K)\Big\{\tilde{f}[\cdot, (t^*_{2k-1})^- - w] - \tilde{f}[\cdot, (t^*_{2k-1})^+ - w] + \tilde{f}[\cdot,(t^*_{2k-1})^-] - \tilde{f}[\cdot,(t^*_{2k-1})^+]\Big\}\delta t_{2k-1} \tag{21}$$

Hence condition (II) of the theorem for this case is

$$\Big\{H[\cdot,t^-_{2k-1} - w] - H[\cdot,t^+_{2k-1} - w] + H[\cdot,t^-_{2k-1}] - H[\cdot,t^+_{2k-1}]\Big\}\delta t_{2k-1} \leqslant 0 \tag{22}$$

The pulse amplitude, as in the case of pulse width modulation, must be determined by a search technique.

SYNTHESIS OF THE OPTIMAL CONTROL

The necessary conditions of the theorem for determining the control amplitudes and switching points of a pulse modulated control, in fact represent terms of the first variation of the performance index with respect to the control. Hence they can be used as gradients in an iteration procedure. The following algorithm is used for finding the optimal control.

(a) Choose an initial control from the admissible set. Integrate the system (5) forward in time from t_0 to the final time t_K.

(b) Knowing the state and the transversality conditions, integrate the costate (7) backward in time from t_K to the initial time t_0.

(c) Knowing the state and costate, iterate the sequences of control amplitudes and switching times according to the rules

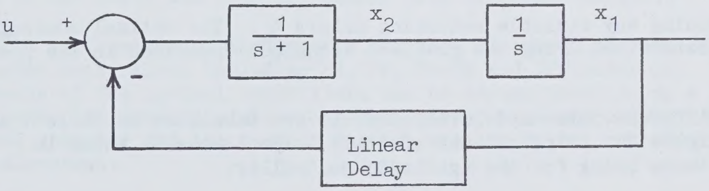
$$\tilde{m}_{k-1}(j+1) = \tilde{m}_{k-1}(j) + C_1 \int_{t_{k-1}(j)}^{t_k(j)} \frac{\partial H}{\partial \tilde{u}}[\cdot,\tilde{m}_{k-1}(j),s]ds \quad , k = 1, \cdots, K \tag{23}$$

$$t_k(j+1) = t_k(j) + C_2\Big\{H[\cdot,\tilde{m}_{k-1}(j),t_k(j)] - H[\cdot,\tilde{m}_k(j),t_k(j)]\Big\}, \quad k = 1, \cdots, K-1 \tag{24}$$

The constants C_1, C_2 are chosen so that the cost function decreases. Using the new control the steps are repeated. The procedure continues until the cost no longer decreases. The final control is then a candidate for the optimal control.

EXAMPLE PROBLEM

Consider the system shown in Figure 1.



The system is described by the equations

$$\begin{aligned}\dot{x}_1(t) &= x_2(t)\\ \dot{x}_2(t) &= -x_1(t - 0.2t) - x_2(t) + u(t)\end{aligned} \tag{25}$$

where the time delay is given by $\theta(t) = 0.2t$ and the initial conditions by

$$x_1(0) = 10, \quad x_2(0) = 1 \tag{26}$$

The control u is to be a form of pulse modulated control with the performance index to be minimized given by

$$J[u] = \frac{1}{2}\int_0^4 [x_1^2(t) + x_2^2(t) + u^2(t)]dt \tag{27}$$

The Hamiltonian, (6), for this problem is

$$H = p_0(x_1^2 + x_2^2 + u^2) + p_1x_2 + p_2(-x_{1\theta} - x_2 + u) \tag{28}$$

The costate equations are obtained from (7).

$$\begin{aligned} &\dot{p}_0(t) = 0, \; t \in [0, 4] \\ &p_0(4) \leqslant 0 \end{aligned} \tag{29}$$

Choose $p_0(t) = -1$; then

$$\begin{aligned} \dot{p}_1(t) &= x_1(t) + 1.25p_2(1.25t), \quad t\varepsilon[0, 3.2] \\ &= x_1(t) \qquad\qquad\qquad\quad\;\, , \quad t\varepsilon[3.2, 4] \\ \dot{p}_2(t) &= x_2(t) - p_1(t) + p_2(t), \quad t\varepsilon[0, 4] \end{aligned} \tag{30}$$

The associated boundary conditions are

$$p_1(4) = 0, \quad p_2(4) = 0 \tag{31}$$

The system performance is to be compared for the various pulse controllers; aperiodic, PA-PW, PA, PW and PPM. Fixed switching points are established at 1 second intervals with appropriate variable switchings between these points. The necessary conditions (I, II) for this problem become,

$$\int_{t_{k-1}}^{t_k} [p_2(t) - u(t)]dt = 0 \tag{32}$$

for determining the pulse amplitude m_k between switching points t_{k-1} and t_k, and

$$H[\cdot, m_{k-1}, t_k^-] = H[\cdot, m_k, t_k^+] \tag{33}$$

for determining any variable switching points t_k. The optimal controls have been determined using the gradient technique described in the previous section.

The different pulse modulated controls are tabulated in Table 1 and shown in Figure 2. A typical set of state trajectories is shown in Figure 3, these being for the aperiodic controller.

For purposes of comparison the optimal continuous control was also found. Note that the necessary conditions for the continuous control are obtained as a special case of the theorem by letting $(t_k - t_{k-1}) \rightarrow 0$. The cost for the continuous control was 78.93, not significantly better than the aperiodic control. Comparing the controllers it is also seen that for this system the easier to implement PA controller may be preferable to the PA-PW and aperiodic controllers if some deterioration in performance is

TABLE I

Pulse Modulator	k	t_{2k-1} sec	m_{2k-2}	m_{2k-1}	Cost
Continuous Control					78.93
Aperiodic	1	0.41	-3.12	0.51	79.48
	2	1.44	2.37	2.69	
	3	2.48	2.15	1.30	
	4	3.34	0.62	0.13	
PA-PW	1	0.36	-3.34	0.00	79.68
	2	2.00	2.65		
	3	2.94	1.80	0.00	
	4	3.44	0.56	0.00	
PA	1		-1.04		81.40
	2		2.77		
	3		1.80		
	4		0.29		
PW $\lvert A\rvert = 1$	1	0.51	-1.00	0.00	84.42
	2	2.00	1.00		
	3	3.00	1.00		
	4	3.35	1.00	0.00	
PPM $\lvert A\rvert = 1$ $w = 0.25$	1	0.26	-1.00		89.86
	2	1.84	1.00		
	3	2.26	1.00		
	4	3.26	1.00		

allowable. The performance of the PW and PPM controllers could be improved by choosing different values for the fixed amplitudes and pulse width.

CONCLUSION

Necessary conditions have been derived for the aperiodically modulated optimal control of systems with a time and state dependent delay in the state vector. (The development has also been extended to the case in which there is a delay in the control vector and to the case in which the components of the state and control vectors have different delays.)

The application of the theorem has been discussed for various pulse modulated controllers including PA, PW, PA-PW and PPM controls. The synthesis of the optimal controllers can be accomplished using a gradient technique. An example problem has been given which was solved using the digital computer and the results for the various pulse controllers have been discussed.

REFERENCES

1. B. Jordan and E. Polak, "Optimal Control of Aperiodic Discrete-Time Systems", J. SIAM Control, 1965, 2, 332.

2. L. S. Pontryagin et al, The Mathematical Theory of Optimal Processes, John Wiley and Sons, Inc., New York, 1962.

3. S. Becker and P. Lopresti, "Continuous Delay Systems under Discrete Control", Proceedings of the Allerton Conference, 1968, 456.

4. J. Banas and A. Vacroux, "Optimal Discrete Control of Time Lag Systems", Proceedings of the IVth IFAC Congress, 1969.

5. D. Kirk, "Optimization of Systems with Pulse-Width Modulated Control", IEEE Trans. A.C., 1967, 12, 307.

6. L. Nardizzi and G. Bekey, "Synthesizing Optimal Controls for Modulated Discrete-Time Systems", Int. J. Control, 1968, 8, 571.

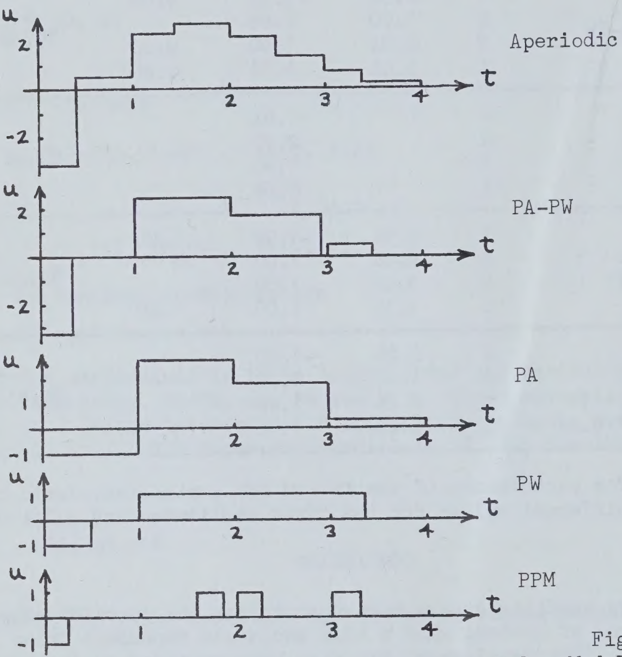


Figure 2
Pulse Modulated Controls

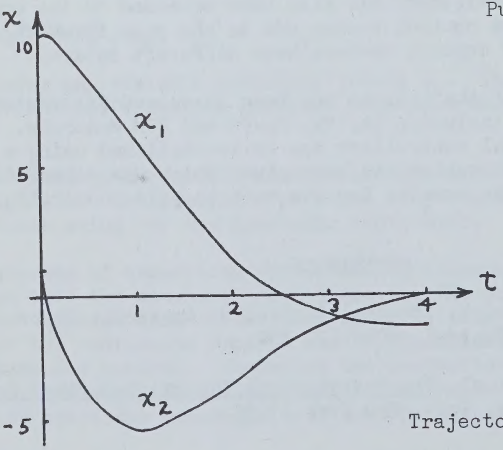


Figure 3
Trajectory for Aperiodic Control

SENSITIVITY REDUCTION IN OPTIMAL CONTROL SYSTEMS

NARESH K. SINHA and SUE-HON DAI
Department of Electrical Engineering
McMaster University
Hamilton, Ontario, Canada

Abstract The implementation of optimal control using two degrees of freedom can be used for the reduction of the integral state sensitivity, while at the same time, either the cost sensitivity or the terminal state sensitivity may be reduced to zero. After the derivation of the general result, a simple example is used for illustration.

INTRODUCTION

An interesting feature of optimal control theory is that while it provides for the determination of the optimal control input to a given plant, it is silent about the method of obtaining this input. This allows considerable freedom in the implementation of optimal control, which may, therefore, be used advantageously in the reduction of the sensitivity of the system to small variations in the parameters of the plant. Considerable work [1,2] has been done on the comparison between open- and closed-loop implementations of optimal control and their effects on the performance and state sensitivities to variations in plant parameters of linear systems with quadratic performance criteria. It is surprising that the implementation with two degrees of freedom has not been considered in the literature on optimal control theory, although, quite a while ago, Horowitz [3] had established the advantages of this configuration which allows the reduction of sensitivity while maintaining the desired overall transfer function.

This paper attempts to show how the configuration with two degrees of freedom may be applied to reduce various sensitivities of an optimal control system to variations in plant parameters.

STATEMENT OF THE PROBLEM

Consider a system described by the vector differential equation

$$\dot{x}(t,\mu) = f(x,u,t,\mu) \tag{1}$$

where $x(t,\mu)$ is the n-dimensional state vector,

$u(t,\mu)$ is the m-dimensional control vector, and

μ is the plant parameter vector, with nominal value μ_o.

The optimal control problem is the determination of the control vector which will minimize the cost function

$$J(t_o,t_f,\mu) = \int_{t_o}^{t_f} L(x,u,t)dt \qquad (2)$$

where t_o, t_f, $x(t_o)=x_o$ and $x(t_f)=x_f$ are all specified, and it is also assumed that μ is in the infinitesimal neighbourhood of μ_o.

The cost sensitivity S_μ^J, the state sensitivity S_μ^x and the terminal sensitivity $S_\mu^{x_f}$ are defined, respectively, as

$$S_\mu^J = \frac{\partial J}{\partial \mu}\bigg|_{\mu=\mu_o} \qquad (3)$$

$$S_\mu^x = y(t) = \frac{\partial x}{\partial \mu}\bigg|_{\mu=\mu_o} \qquad (4)$$

and

$$S_\mu^{x_f} = y(t_f) = \frac{\partial x_f}{\partial \mu}\bigg|_{\mu=\mu_o} \qquad (5)$$

All these sensitivities are functions of an implementation vector, v(t), defined as

$$v(t) = \frac{\partial u(t,\mu)}{\partial \mu}\bigg|_{\mu_o} \qquad (6)$$

The optimal control input for the nominal case ($\mu=\mu_o$), denoted by $u(t,\mu_o) = u^*(t)$, can be determined by applying Pontryagin's minimum principle, i.e., by obtaining $u^*(t)$ such that the Hamiltonian, H, is minimized, where

$$H = L(x,u,t) + f^T(x,u,t,\mu_o) \cdot p(t) \qquad (7)$$

and

$$\dot{p}(t) = -\frac{\partial H}{\partial x} \qquad (8)$$

where p(t) is the nominal costate vector.

This optimal control input may be implemented by using a combination of open- and closed-loop formulations, leading to the system with two degrees of freedom. In this case one may write

$$u(t,\mu) = M(t)\ x(t,\mu) + g(t) \qquad (9)$$

where M(t) is the matrix representing state-variable feedback and g(t) represents the open-loop portion of the control input.

Since $u(t,\mu)$ must be optimal for the nominal case ($\mu=\mu_o$), the following equation must be satisfied by all admissible M(t) and g(t),

$$u(t,\mu_o) = u^*(t) = M(t)\ x^*(t) + g(t) \qquad (10)$$

where $x^*(t)$ is the optimal state trajectory for the nominal case, i.e.,

$$x^*(t) = x(t,\mu_o)\Big|_{u=u^*(t)} \tag{11}$$

In particular, when equations (1) and (2) represent a linear system with a quadratic cost function, it is possible to make either $M(t) = 0$ or $g(t) = 0$, and still make the control optimal for the nominal case. These are the so-called "open-loop" and "closed-loop" implementations of optimal control which have been investigated by many authors and compared from the point of view of sensitivity to variations in plant parameters. It appears natural that the implementation represented by equation (2) provides more flexibility which could be utilized for the reduction of the various sensitivities defined in equations (3), (4) and (5).

Thus, the problem may be stated as the selection of $M(t)$ in such a manner that these sensitivities are minimized. However, a scalar measure for each of the sensitivities must be defined in order that this minimization may be carried out. In the case where μ is a scalar, S_μ^J is also a scalar. But S_μ^x is a time-varying vector, and one may define the "integral" state sensitivity as

$$S_y = \int_{t_o}^{t_f} y^T(t) R\, y(t)\, dt \tag{12}$$

where R is a given symmetric positive definite matrix. The terminal sensitivity $S_\mu^{x_f}$ is a vector independent of time, and thus its norm may be used as a measure of this sensitivity. If μ is a vector, scalar measures can be obtained, in each case, by defining suitable norms for the various sensitivity vectors or matrices.

INSENSITIVE DESIGNS WITH INTEGRAL STATE SENSITIVITY REDUCTION

Following Kokotovic and Heller [4], it can be shown that the cost sensitivity (for the case where μ is a scalar) is given by

$$S_\mu^J = -y^T\, p^* \Big|_{t_o}^{t_f} + \int_{t_o}^{t_f} \frac{\partial f}{\partial \mu}\Big|_{\mu=\mu_o}^T p^* dt \tag{13}$$

Thus, the cost sensitivity can be made zero if

$$y^T\, p^* \Big|_{t_o}^{t_f} = \int_{t_o}^{t_f} \frac{\partial f}{\partial \mu}\Big|_{\mu=\mu_o}^T p^* dt \tag{14}$$

In general, the initial state is independent of μ, and hence

$$y(t_o) = 0. \tag{15}$$

Thus the condition for cost insensitivity may be expressed as

$$y^T(t_f) \cdot p(t_f) - \int_{t_o}^{t_f} \frac{\partial f}{\partial \mu}\bigg|_{\mu=\mu_o}^{T} p^* dt \tag{16}$$

The state sensitivity vector y(t) is obtained by differentiating (1) partially with respect to μ. Hence,

$$\dot{y}(t) = \frac{\partial f}{\partial x}\bigg|_{\mu=\mu_o} y + \frac{\partial f}{\partial u}\bigg|_{\mu=\mu_o} v + \frac{\partial f}{\partial \mu}\bigg|_{\mu=\mu_o} \tag{17}$$

The implementation vector v(t) is obtained by differentiating (9) partially with respect to μ. Hence,

$$v(t) = M\ y \tag{18}$$

Combining equations (17) and (18), we have

$$\dot{y}(t) = h(t,y,M) = \{ \frac{\partial f}{\partial x} + \frac{\partial f}{\partial u} M\}_{\mu=\mu_o} y + \frac{\partial f}{\partial \mu}\bigg|_{\mu=\mu_o} \tag{19}$$

The functions $\frac{\partial f}{\partial x}\Big|_{\mu=\mu_o}$, $\frac{\partial f}{\partial u}\Big|_{\mu=\mu_o}$ and $\frac{\partial f}{\partial \mu}\Big|_{\mu=\mu_o}$ are assumed to be known and depend upon x*(t) and u*(t) which can be considered as known functions of time. If the components of h, $\frac{\partial h}{\partial t}$ and $\frac{\partial h}{\partial y}$, are continuous on $R^n\ x(t_o,t_f)$ and $R^{nxq}\ x(t_o,t_f)$, then an optimal implementation problem can be formulated as follows: to find M(t) driving $y(t_o)=0$ to a given point or set $y(t_f)$ via (19) and minimizing the integral state sensitivity S_y defined in (12). In the terminal insensitive design $y(t_f)$ is equated to a zero vector, whereas in the cost insensitive design, $y(t_f)$ is a point on the hyperplane defined by (16). In general, it is seldom possible to obtain cost insensitivity and terminal insensitivity simultaneously. By making $y(t_f)$ a point on the hyperplane defined by (16) at a minimum distance from the origin, one may achieve cost insensitivity, while minimizing terminal sensitivity and the integral state sensitivity.

The necessary condition for the optimal implementation M*(t) can be obtained by applying Pontryagin's minimum principle to equations (19) and (12). In addition, from practical considerations, the integral of the norm-squared of M(t), i.e., $\int_{t_o}^{t_f} ||M(t)||^2 dt$ must be finite. If desired, a more restrictive magnitude constraint on ||M(t)|| may also be imposed.

EXAMPLE

The method will now be applied to a second-order linear system with one variable parameter. Consider a system described by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1+\mu \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \tag{20}$$

where $\mu_o = 0$ for the nominal case. The cost function is given by

$$J = \int_o^\infty (x_1^2 + x_2^2 + \frac{1}{4}u^2)\,dt \tag{21}$$

with the initial and terminal states being given as

$$x_o = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad x_f = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{22}$$

Application of Pontryagin's minimum principle to the nominal case gives

$$x_1^*(t) = 2e^{-t} - e^{-t} \tag{23}$$

$$x_2^*(t) = -2e^{-t} + 2e^{-2t} \tag{24}$$

$$p_1^*(t) = 4e^{-t} - e^{-2t} \tag{25}$$

$$p_2^*(t) = e^{-2t} \tag{26}$$

and $$u^*(t) = -2e^{-2t} \tag{27}$$

The determination of the matrix

$$M = [m_1, m_2] \tag{28}$$

will now be considered.

From (20), the following partial derivatives are obtained

$$\frac{\partial f}{\partial x} = \begin{bmatrix} 0 & 1 \\ 0 & -1+\mu \end{bmatrix} \tag{29}$$

$$\frac{\partial f}{\partial u} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{30}$$

$$\frac{\partial f}{\partial \mu} = \begin{bmatrix} 0 \\ x_2 \end{bmatrix} \tag{31}$$

Using these in equation (19), the following differential equations are used for the state sensitivity,

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ m_1 & -1+m_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ x_2 \end{bmatrix} \tag{32}$$

The problem now is to determine $m_1(t)$ and $m_2(t)$ which will minimize

$$S_y = \int_0^\infty (y_1^2 + y_2^2)dt \tag{33}$$

while satisfying the constraints

$$y(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$y(\infty) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{34}$$

and $\int_0^\infty (m_1^2 + m_2^2)dt < \infty$

The solution of this problem provides M(t), which can be substituted in (9) to obtain g(t).

Thus, it will be seen that the implementation using two degrees of freedom has been used to minimize the integral state sensitivity and also reduce the terminal sensitivity to zero.

CONCLUSION

It has been shown that the implementation of optimal control with two degrees of freedom can be used for the reduction of integral state sensitivity and also to make either the cost or the terminal sensitivity zero in the case of systems with one variable parameter. The extension to the multiparameter sensitivity problem is currently under investigation.

REFERENCES

[1] D.C. Youla and P. Dorato, "On the comparison of sensitivities of open-loop and closed-loop optimal control systems", IEEE Trans. Automatic Control, (Short Papers), Vol. AC-13, pp. 186-188, April 1968.

[2] E. Kriendler, "Closed-loop sensitivity reduction of linear optimal control systems", IEEE Trans. Automatic Control, Vol. AC-13, pp. 254-262, June 1968.

[3] I.M. Horowitz, "Fundamental Theory of Automatic Linear Feedback Control Systems", IRE Trans. Automatic Control, Vol. AC-4, pp. 5-19, 1959.

[4] P. Kokotovic and J. Heller, "Direct and adjoint sensitivity equations for parameter optimization", IEEE Trans. Automatic Control (Correspondence), Vol. AC-12, pp. 609-610, October 1967.

LIMIT CYCLES IN INSTRUMENT SERVOS

K. W. HAN
Chung-Shan Institute and
National Chiao-Tung University
Hsinchu, Taiwan

G. J. THALER
Naval Postgraduate School
Monterey, California

SUMMARY

Instrument servos are often second order systems designed for high performance. The geartrains used may have some backlash, and amplifiers are often driven into saturation so that several nonlinearities exist. If the servo is lightly damped a limit cycle may result due to the backlash and saturation. Analysis of such systems may be undertaken on the phase plane, but standard techniques are unnecessarily laborious. In this paper a new concept called a "critical line" is introduced.

The "critical line" is defined as the locus of all points on the phase plane such that the x-coordinate of each point is the negative of the x-coordinate of the horizontal axis intercept of the phase trajectory passing through the point. For a linear system the critical line is a straight line with slope determined by the damping ratio. When a servo is saturated the critical line is curved, as it is for the general case of a nonlinear plant.

Limit cycles due to backlash are symmetrical on the phase plane and, therefore, must have one terminal point on the critical line. It follows that the limit cycle amplitude is defined by the intersection of the critical curve with the backlash recombination line. This is verified by computer simulation.

Saturation in the servo amplifier reshapes both the critical curve and the backlash recombination line. The amplitude of the limit cycle tends to increase, as predicted by the intersection of critical curve and recombination line.

When tachometer feedback is used the linear zone on the phase plane is rotated and the critical curve is reshaped as is the backlash recombination line. For some conditions three limit cycles (two stable, one unstable) are shown to be possible. Simulation results confirm the predictions.

CONSTRUCTION OF CRITICAL CURVES

The terminology "critical curve" used in this paper means that in a phase-plane, if a trajectory passes through a point with coordinate $(-\alpha\ ,\ \omega)$ on the critical curve, then the nearest intersection of the trajectory with the real axis of the phase plane is at α.

In order to construct a critical curve, a phase-portrait of a system with no backlash is used. The phase-portrait can be obtained by calculation, by a phase-plane plot, or by an analog computer. A typical phase-portrait is given in Fig. 1, and the procedure of constructing a critical curve is as follows:

(1) On the real axis of the phase-plane measure the distance between the origin (0) and a point (A) where a trajectory crosses the real axis, and then mark the same distance OB on the opposite side of the real axis as indicated in Fig. 1.

(2) From point B draw a line parallel to the vertical axis, and find its intersection (P) with the trajectory which passes through point A, then P is a point on the critical curve.

(3) For various values of OA, using the same method, a set of points (as point P) can be found, which defines a critical curve.

From the general shape of the phase-portrait, it can be seen (or proved easily) that the critical curve in Fig. 1 is a straight line. For various values of damping ratio (ζ), the angles between the critical lines and the negative real axis are represented by a curve as in Fig. 2, which can be used to find the critical curves without a phase-portrait.

It is interesting to note that if a trajectory jumps vertically from a point ($-\alpha$) on the real axis of the phase-plane to a point on the critical curve, then the system has a limit cycle with amplitude α.

Note also that the proposed procedure can be used to construct the critical curves of any linear or nonlinear systems (2nd order) as long as a phase-portrait can be obtained.

CONSTRUCTION OF RECOMBINATION CURVES

This section presents a simplified method for finding the recombination curves of 2nd order linear systems with simple backlash.

Consider the system in Fig. 3, if the state point is in the backlash region, then the motor velocity (θ_m) is

$$\frac{\theta_m}{w_n} = \frac{E_b(1-e^{-2\zeta\omega_n t})}{2\zeta} \qquad (1)$$

where E_b (a constant) is the error of the system when the state point is in the backlash region. Assume the system has a backlash Δ, then the time (t_r) required to take up the backlash is defined by

$$\begin{aligned} \Delta &= \int_o^{t_r} \theta_m \, dt \\ &= \frac{\omega_n E_b}{2\zeta} \left[t_r + \frac{1}{2\zeta\omega_n} \left(e^{-2\zeta\omega_n t_r} - 1\right)\right] \qquad (2) \end{aligned}$$

Let $P = 2\zeta\omega_n$ and $K = \omega_n^2$, then Eq. (2) becomes

$$\frac{\Delta P^2}{KE_b} = Pt_r + (e^{-Pt_r} - 1) \qquad (3)$$

Let $H = P^2\Delta/KE_b$ and $X = Pt_r$, then

$$H = X + e^{-X} - 1 \qquad (4)$$

For various values of X, a curve can be plotted as in Fig. 4. If the values of ζ, Δ and K are given, for each value of E_b the value of H can be calculated, and X can be read out from Fig. 4; thus the rest time (t_r) can be found.[1]

From Eq. (1), let $H' = E_b/2\zeta$ and $\eta = 1-e^{-X}$, then

$$\frac{E}{H'\omega_n} = \eta = 1-e^{-X} \qquad (5)$$

STABILITY CRITERION

After the critical curve and the recombination curve of a system are found, a stability criterion can be defined as follows:

In the second quadrant of the phase-plane, if the critical curve is above (below) the recombination curve, then the system is stable (unstable); and if there is an intersection between these two curves, then the system has a limit cycle.

The reason for this stability criterion is quite evident, since the effect of simple backlash is to make each of the trajectories have a vertical jump from the real axis, and the maximum amount of jump is defined by the recombination curve.

Example 1. Consider the system in Fig. 3, if $\zeta = 0.2$, $K = 1$ and $\Delta = 0.05$, then

$P = 2\zeta\omega_n = 0.4 \qquad H = 0.008/E_b \qquad H' = E_b/0.4$

For various values of E_b, using Fig. 4, the following table can be obtained.

1	E_b	0.5	0.3	0.2	0.1	0.05
2	H	0.016	0.126	0.04	0.08	0.16
3	H'	1.25	0.75	0.5	0.25	0.125
4	X	0.2	0.26	0.31	0.43	0.62
5	$\frac{E}{H'\omega_n}$	0.18	0.235	0.27	0.35	0.46
6	E/ω_n	0.225	0.176	0.135	0.0875	0.0575
7	t_r	0.50	0.65	0.775	1.075	1.55

Using the numerical values in the first and the sixth rows, a recombination curve can be drawn as in Fig. 5, where a critical curve for $\zeta = 0.2$ is drawn also. The intersection point (Q_1) indicates that the system has a stable limit cycle at $E_b = 0.032$.

The result in this example has been checked with that in reference 1. The time required for calculation is substantially reduced.

SATURATED SYSTEM WITH BACKLASH

In this section, a 2nd order system with two nonlinearities, saturation and backlash, is analyzed. The block diagram of the system is in Fig. 6, and its characteristics can be classified into four cases as represented by the following differential equations:

(1) Linear case; i.e., no saturation and not in the backlash region

$$\theta_c + P\theta_c = KE = K(\theta_r + \theta_c) \qquad (6)$$

(2) Saturated but not in the backlash region

$$\theta_c + P\theta_c = KE_s \qquad (7)$$

(3) In the backlash region but not saturated

$$\theta_m + P\theta_m = KE_b \tag{8}$$

(4) Saturated and in the backlash region

$$\theta_m + P\theta_m = KE_s \tag{9}$$

where E_s in the value of the error (E) which just saturates the main amplifier; thus the linear region in the phase-plane is defined by two vertical lines ($E = \pm E_s$), and inside the linear region the methods proposed can be used to find the recombination curve and the critical curve.

In the saturated region, both the critical curve and the recombination curve are changed. To find the critical curve, a phase-portrait of the saturated system (without backlash) is required. For $K = 50$, $\zeta = 0.282$ and $E_s = 0.04$, a phase-portrait is plotted in Fig. 7a, where the critical curve is indicated.

In order to construct the recombination curve, the motor velocity in the backlash region is required. If the system is under saturated condition, then

$$\theta_m(t) = \frac{KE_s}{P}(1-e^{-pt}) \tag{10}$$

and the maximum value of $\theta_m(t)$ is

$$\theta_m(\infty) = KE_s/P \tag{11}$$

If $E_b > E_s$, and the backlash region is wide enough to let the motor reach its maximum velocity, then the recombination curve is a straight line parallel to the real axis. Thus, as indicated in Fig. 7a, a recombination curve may consist of three sections defined by

$$\theta_m(t) = \frac{KE_b}{P}(1-e^{-pt}) \qquad \text{(OM)} \tag{12}$$

$$\theta_m(t) = \frac{KE_s}{P}(1-e^{-pt}) \qquad \text{(MR)} \tag{13}$$

and $$\theta_m(\infty) = \frac{KE_s}{P} \qquad \text{(RW)} \tag{14}$$

After the critical curve and the recombination curve are found, the characteristics of the system can be defined, but some modifications about the limit cycle must be made as follows:

In Fig. 7a, although the intersection point (P) defines a limit cycle, the magnitude of the limit cycle is reduced, the maximum velocity of the system is limited due to saturation. In other words, the actual size of the limit cycle is decided by the width of the linear zone, the maximum velocity of the motor, and the damping characteristics of the system.

Fig. 7a also indicates that, if the system is linear, the magnitude of the limit cycle is decided by point Q_1. A computer simulation gives these two limit cycles as in Figs. 7b and 7c respectively.

For the considered system, if there is no limit cycle in the unsaturated case, then the system will have no limit cycle in the saturated case, since the effective damping of the system is increased due to saturation.

THE EFFECT OF TACHOMETER FEEDBACK

For a linear 2nd order system, the phase-portrait is decided by the damping ratio alone regardless the method used for damping, but for a saturated system the results are quite different.

Consider the system in Fig. 8, for small values of E and ϵ the system is linear; thus the method previously proposed can be applied. For large E, if the system is not saturated (i.e. $\epsilon < \epsilon_s$), then the system equation is

$$\theta_m(t) = K\epsilon(t) = K\,[E-K_T\theta_m(t)] \tag{15}$$

If saturated, then

$$\theta_m(t) = K\,\epsilon_s \tag{16}$$

where ϵ_s is the minimum input of the main amplifier to make it saturate. From Fig. 8, it can be seen that the effect of tachometer feedback is to form a minor loop which makes the maximum velocity of the motor, in the backlash region, equal to

$$\theta_m(t)_{max.} = E_b/K_T \tag{17}$$

If the system is saturated, then the motor is driven by a constant input (ϵ_s); thus Eq. (16) gives

$$\theta_m(t) = Kt\epsilon_s \tag{18}$$

From Eq. (17), the maximum motor velocity is inversely proportional to the tachometer feedback; thus the recombination curve may coincide with the center line of the linear zone.

In case Δ is small, and the motor velocity in the backlash region can not reach its maximum value, then the time for taking up the backlash is defined by

$$\Delta = \int_0^{t_r} Kt\epsilon_s dt = \tfrac{1}{2}K\epsilon_s t_r^2 \tag{19}$$

which gives

$$t_r = (2\Delta/K\epsilon_s)^{1/2} \tag{20}$$

Thus the motor velocity at the recombination instant is

$$\theta_{mr} = Kt_r\epsilon_s = (2K\epsilon_s\Delta)^{1/2} \tag{21}$$

Finally, using Eqs. (18) and (21), for specified value of Δ, the recombination curve can be constructed.

Since the slope of the linear zone is inversely proportional to K_T, to increase K_T may increase the effective damping of the system; but on the other hand, to incrase K_T will reduce the width of the linear zone, which will reduce effective damping. The application of the critical curve method for analyzing such a system is illustrated by the following example.

Example 2. Consider the system in Fig. 8, if $\Delta = 0.1$, $K = 50$, $K_T = 0.0565$ and $\epsilon_s = 0.04$, then the isocline equation for the saturated region is

$$E/\omega_n = -0.04/N \qquad \text{where } N = \frac{dE/\omega_n}{dE}$$

A phase-portrait and a critical curve are plotted as in Fig. 9. After the recombination curve is constructed, the result indicates that there is a large limit cycle at E = 0.2, which has been checked with an analog computer.

If the main amplifier of the system is unsaturated, a limit cycle is found at E = 0.081; thus the effect of saturation has reduced the effective damping of the system.

For the considered system, there are four important parameters Δ, ζ, ϵ_s and K_T. For various combinations of these parameters, one can obtain many interesting results. For example, if K_T is adjusted to give $\zeta = 0.27$, a very small limit cycle will exist inside the unsaturated zone; in addition there is a large stable limit cycle together with an unstable limit cycle as indicated in Fig. 10. All three limit cycles have been checked using an analog computer.

Since control systems with saturation, backlash, and rate feedback are commonly used in aircrafts and missiles, the method proposed in this section might be useful for analysis and design of such control systems.

CONCLUSIONS

In this paper, a critical curve method has been presented, which can be used to analyze 2nd order systems with two nonlinearities. For a system with simple backlash, the recombination curve and the rest time can be found easily. The effects of saturation and tachometer feedback have been analyzed, and a system with three limit cycles has been illustrated.

REFERENCES

[1] G. J. Thaler and R. G. Brown, "Analysis and Design of Feedback Control Systems," McGraw-Hill Book Company, New York, 1969, PP 462-472.

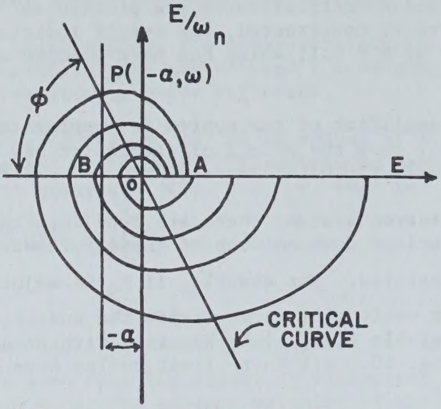
[2] H. H. Chuang, "Analysis of 2nd Order System with Simple Backlash," M.S. Thesis, EE Dept., Cheng-Kung University, Taiwan, 1963.

[3] C. Y. Chiang, "Analysis of Nonlinear Control Systems Using Critical Method," M. S. Thesis, Institute of Electronics, Chiao-tung University, Taiwan, 1964.

[4] W. J. Lutkenhouse, W. J.: "Dividing Lines for Backlash in the Phase Plane," M. S. Thesis, U.S. Naval Postgraduate School, Monterey, Calif., 1959.

[5] M. P. Pastel and G. J. Thaler: Instrument Servomechanisms with Backlash, Coulomb Friction and Stiction, Trans. AIEE, 1960.

[6] G. J. Thaler and M. P. Pastel, "Analysis and Design of Nonlinear Feedback Control Systems", McGraw Hill Book Company, New York, 1962.



Ė/ωn
P(-α,ω)
φ
B
A
O
E
-α
CRITICAL CURVE

FIG. I PHASE PORTRAIT OF A SECOND ORDER LINEAR SYSTEM

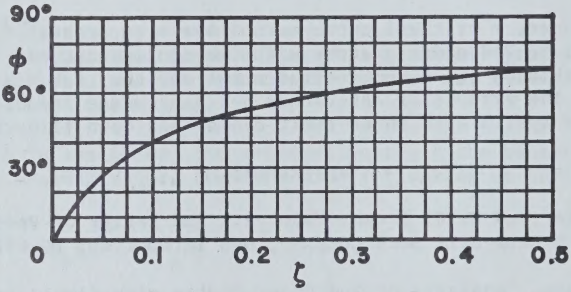
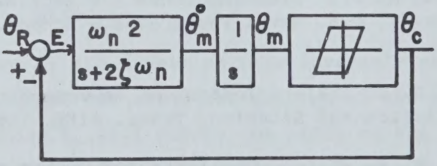


FIG. 2 RELATIONSHIP BETWEEN SLOPE OF THE CRITICAL CURVE AND THE DAMPING RATIO



θR
E
+
−
ωn 2
s+2ζωn
θ̇m
1
s
θm
θc

FIG. 3 BLOCK DIAGRAM, SECOND ORDER SYSTEM WITH BACKLASH

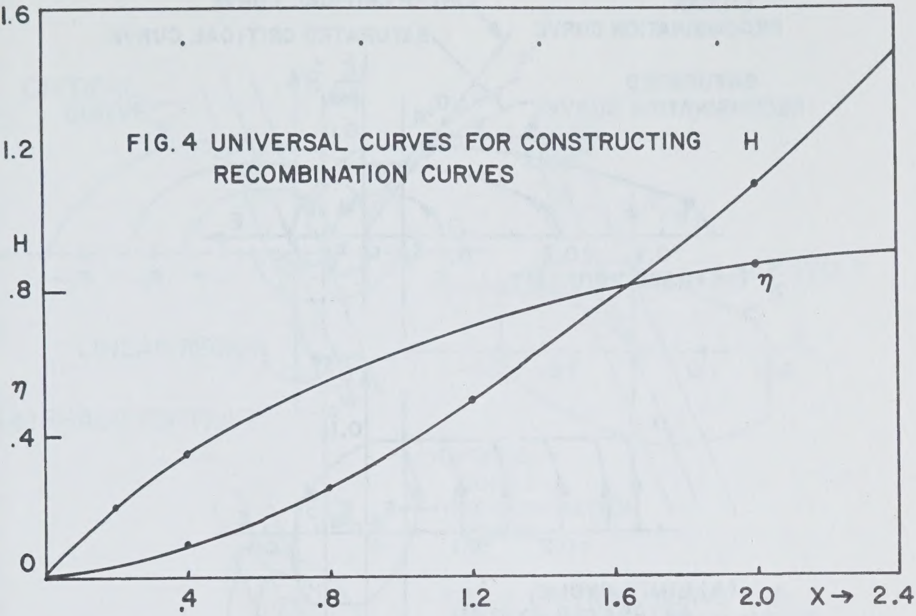


FIG. 4 UNIVERSAL CURVES FOR CONSTRUCTING RECOMBINATION CURVES

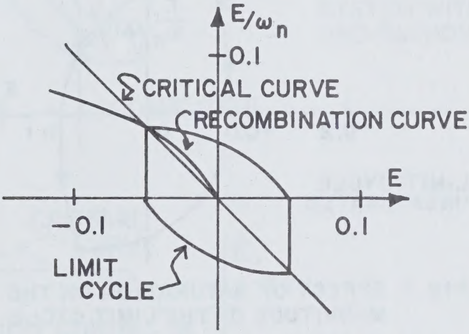

$\dot{E}/\omega_n$
0.1
CRITICAL CURVE
RECOMBINATION CURVE
E
−0.1
0.1
LIMIT CYCLE

FIG. 5 DEFINITION OF A LIMIT CYCLE AT INTERSECTION OF CURVES

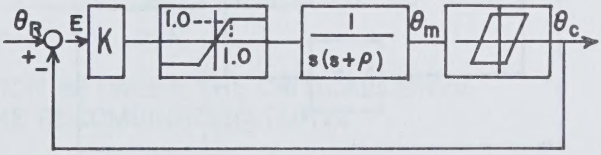
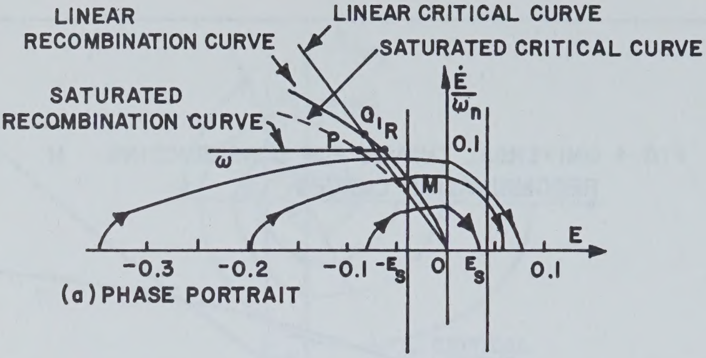
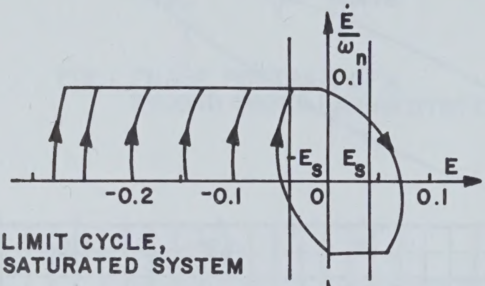

θ_R
E
+
−
K
1.0
1.0
$\frac{1}{s(s+\rho)}$
θ_m
θ_c

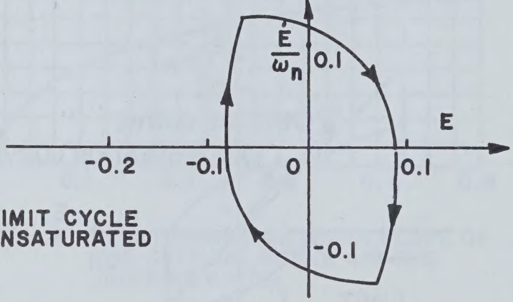
FIG. 6 BLOCK DIAGRAM OF A SATURATING SECOND ORDER SYSTEM WITH BACKLASH



(a) PHASE PORTRAIT



(b) LIMIT CYCLE, SATURATED SYSTEM



(c) LIMIT CYCLE UNSATURATED

FIG.7 EFFECT OF SATURATION ON THE MAGNITUDE OF THE LIMIT CYCLE

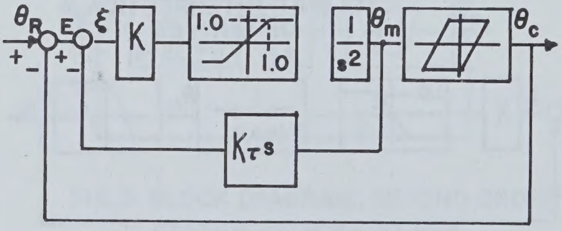
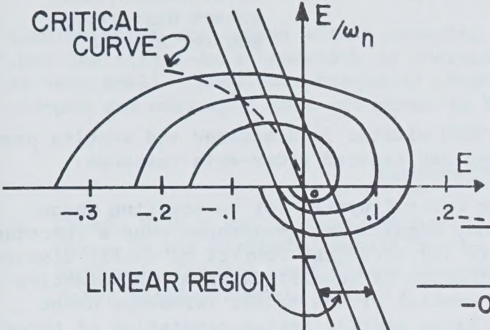

θ_R
+
-
E
+
-
ξ
K
1.0
1.0
1/s²
θ_m
θ_c
K_τ s

FIG.8 SERVO WITH SATURATION, BACKLASH AND TACHOMETER FEEDBACK


CRITICAL CURVE
Ė/ωn
E
−.3
−.2
−.1
.1
.2
LINEAR REGION

(a) PHASE PORTRAIT

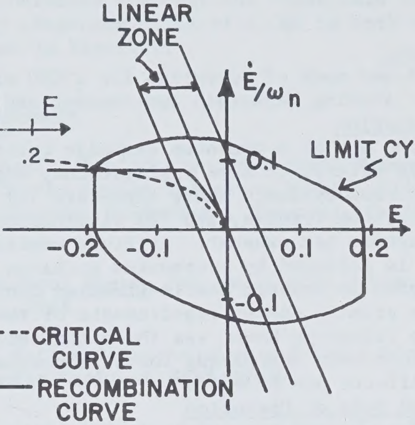
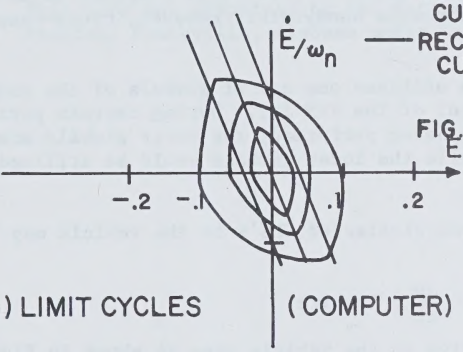
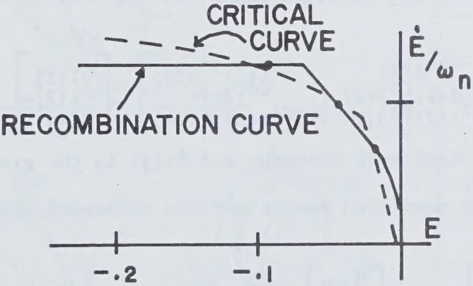

LINEAR ZONE
Ė/ωn
LIMIT CYCLE
0.1
E
−0.2
−0.1
0.1
0.2
−0.1
----CRITICAL CURVE
——RECOMBINATION CURVE

FIG. 9 CRITICAL CURVE AND RECOMBINATION CURVE FOR THE SATURATED SYSTEM WITH BACKLASH AND TACHOMETER FEEDBACK



(b) LIMIT CYCLES (COMPUTER)


CRITICAL CURVE
Ė/ωn
RECOMBINATION CURVE
E
−.2
−.1

(c) RELATION BETWEEN THE CRITICAL CURVE AND THE RECOMBINATION CURVE

FIG. 10 SYSTEM WITH THREE LIMIT CYCLES

ON A CLAMPED ATTITUDE-CONTROL MODE FOR AN ORBITING SPACE VEHICLE[*]

DONALD W. SUTHERLIN
Auburn University
Auburn, Alabama

J. S. BOLAND, III
Auburn University
Auburn, Alabama

Abstract

A new mode of operation for a CMG cluster is discussed and results presented showing bandwidth improvement and reduced cross-axis response.

Introduction

The use of a momentum exchange control system for an orbiting space vehicle offers, in some applications, significant advantages over a reaction jet control system. Among these are (1) efficient control of cyclic disturbance system torques, and (2) elimination of possible system contamination by reaction jet exhaust. Attitude control of the Apollo Telescope Mount (ATM) is achieved by a momentum exchange control system consisting of three orthogonally-mounted double-gimbaled Control Moment Gryos (CMG's) [1]. One of the primary design requirements of the attitude control system for the Apollo Telescope Mount was the minimization of cross-axis response while simultaneously maximizing the direct-axis bandwidth. However, cross-coupling effects are present.

Clamped Mode of Operation

The clamped mode of operation utilizes one set of gimbals of the double-gimbaled CMG's for attitude control of the ATM [2]. During certain periods, such as when solar experiments are being performed, the outer gimbals would be rigidly clamped to the frame while the inner gimbals would be utilized for dynamic control.

Optimal Steering Law

The reaction moment of an ideal cluster of CMG's on the vehicle may be expressed as:

$$M_{RV} = -\frac{dH}{dt}\Big|\ \text{I-space} \tag{1}$$

With the CMG cluster located relative to the vehicle axes as shown in Figure 1 and the CMG's in the clamped mode, the relationship between the 3 inner gimbal rates of the CMG cluster and the reaction moment on the vehicle axes may be expressed as:

$$\begin{bmatrix} M_{RXV} \\ M_{RYV} \\ M_{RZV} \end{bmatrix} = H \begin{bmatrix} S\delta_{1(1)}C\delta_{3(1)} & C\delta_{1(2)} & -S\delta_{1(3)}S\delta_{3(3)} \\ -S\delta_{1(1)}C\delta_{3(1)} & S\delta_{1(2)}C\delta_{3(2)} & C\delta_{1(3)} \\ C\delta_{1(1)} & -S\delta_{1(2)}C\delta_{3(2)} & S\delta_{1(3)}C\delta_{3(3)} \end{bmatrix} \begin{bmatrix} \dot{\delta}_{1(1)} \\ \dot{\delta}_{1(2)} \\ \dot{\delta}_{1(3)} \end{bmatrix} \tag{2}$$

where H is the magnitude of the wheel spin momentum and $\dot{\delta}i(j)$ is the gimbal rate of i^th^ pivot of the j^th^ CMG.

The relationship between the commanded moment and the commanded gimbal rate may be expressed as

$$\begin{bmatrix} \dot{\delta}_{1(1)C} \\ \dot{\delta}_{1(2)C} \\ \dot{\delta}_{1(3)C} \end{bmatrix} = [T_s] \begin{bmatrix} M_{CXV} \\ M_{CYV} \\ M_{CZV} \end{bmatrix} \tag{3}$$

where $[T_s]$ is the steering law. Assuming $\dot{\delta}i(j) = \dot{\delta}i(j)C$, the expression between the commanded moment and the reaction moment may be expressed as:

$$\begin{bmatrix} M_{RXV} \\ M_{RYV} \\ M_{RZV} \end{bmatrix} = [A]\ [T_s] \begin{bmatrix} M_{CXV} \\ M_{CYV} \\ M_{CXV} \end{bmatrix} \tag{4}$$

In order to optimally control the CMG cluster, it is necessary that

$$[A]\,[T_s] = [I] \tag{5}$$

Simulation of the system with a steering law such that $[T_s] = [A]^{-1}$ shows that the direct-axis bandwidth is increased and that the cross-axis response is very small. Amplitude-frequency characteristics of a CMG in both the clamped and unclamped mode are shown in Figure 2.

REFERENCES

1. Chubb, W. B.; and Epstein, Michael: Application of Control Moment Gyros in the Attitude Control of the Apollo Telescope Mount. AIAA Guidance, Control, and Flight Dynamics Conference, Pasadena, Calif., paper 68-866, August, 1968.

2. Borelli, M. T.: An Improved Control System for the Apollo Telescope Mount. NASA TN X-53830, April, 1969.

[*] This work was supported by the National Aeronautics and Space Administration, Huntsville, Alabama under Contract NAS8-20104.

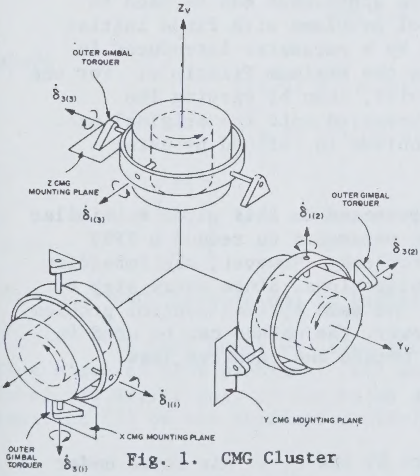

Z_V
OUTER GIMBAL TORQUER
$\dot{\delta}_{3(3)}$
Z CMG MOUNTING PLANE
$\dot{\delta}_{1(3)}$
OUTER GIMBAL TORQUER
$\dot{\delta}_{1(2)}$
$\dot{\delta}_{3(2)}$
Y_V
$\dot{\delta}_{1(1)}$
Y CMG MOUNTING PLANE
X CMG MOUNTING PLANE
OUTER GIMBAL TORQUER
$\dot{\delta}_{3(1)}$

Fig. 1. CMG Cluster

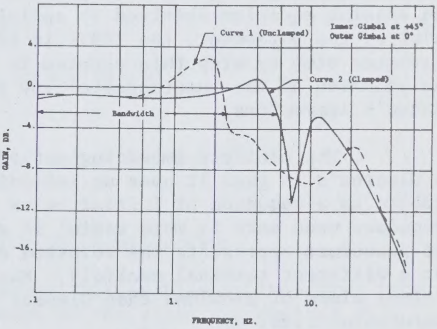


FIG. 2. AMPLITUDE-FREQUENCY CHARACTERISTICS

A MANIFOLD-IMBEDDING SOLUTION OF OPTIMAL CONTROL PROBLEMS*

G. D'ANS, Y. HONTOIR, M. JAMSHIDI
Department of Electrical Engineering
and Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

ABSTRACT

A new imbedding method is developed for solving two-point boundary value (TPBV) problems arising in optimal control theory. The TPBV problem is solved by varying a parameter introduced in the boundary conditions. Newton's iterations are combined with the imbedding method to refine the solution. The method is applied to a third-order nonlinear plant.

I. INTRODUCTION

Two types of computational methods are commonly used for the solution of two-point boundary value (TPBV) problems appearing in optimal control theory. First are the "iterative" methods such as gradient's and Newton's [1] where a choice of initial guess and convergence properties are among the basic difficulties. Second are the "imbedding" methods such as dynamic programming [2] and the invariant imbedding [3] where the solution of the original TPBV problem is obtained by solving a "class" of similar problems. Dynamic programming requires a prohibitive amount of computation and memory storage while the invariant imbedding leads to the solution of linear partial differential equations.

A parameter variation method recently developed by Glasser [4] combining both the imbedding and iterative approaches can be used to treat specified final time optimal control problems with fixed initial state. The imbedding class is generated by a parameter introduced in the adjoint equation obtained by applying the Maximum Principle. For one value of the parameter, the TPBVP is trivial, then by varying the parameter step by step this problem is converted into the original one. The solution of the intermediate TPBV problems is refined by using Newton's iterations.

The manifold-imbedding method proposed in this paper is similar to Glasser's in that it uses an imbedding parameter to reduce a TPBV problem to a sequence of initial value problems. However, the imbedding procedure used here is more useful in applications, since every step of the procedure represents the solution of the same optimal control problem for a different terminal manifold. Moreover, the method can be used in a wider class of problems than Glasser's method and requires less computation time.

---

*This work was supported in part by the U. S. Air Force under Grant AFOSR-68-1579C, in part by the Joint Services Electronics Program (U. S. Army, U. S. Navy, and U. S. Air Force) under Contract DAAB-07-67-C-0199, and in part by the National Science Foundation under Grant GK-3893, Coordinated Science Laboratory, University of Illinois

II. THE MANIFOLD IMBEDDING METHOD

Consider a system

$$\dot{x} = \Psi(x,u) \qquad x(0) = x_o \subset R^n \tag{1}$$

with the cost functional

$$J = \int_o^T L(x,u)dt \tag{2}$$

and control $u \subset R^s$. Functions Ψ and L are C^1 and the state $x(T)$ is required to lie on the manifold τ,

$$\tau = \{x \in R^n : w(x) = 0,\ w(x) \in R^m\} \tag{3}$$

where $w \in C^2$ and for $x \in \tau$, rank $w_x(x) = m$.

The problem is to find a control $u^*(t)$, $t \in [0,T]$ steering the system for x_o to τ in an unspecified time T so that J is minimized. Using the necessary conditions of the maximum principle, the following TPBV problem is obtained,

$$\dot{x} = f(x,p) \tag{4}$$

$$\dot{p} = g(x,p) \tag{5}$$

$$x(0) = x_o \tag{6}$$

$$w[x(T)] = 0 \tag{7}$$

$$p(T) = w_x'[x(T)]\mu \tag{8}$$

$$M[x_o,p(0)] = 0 \tag{9}$$

where

$$f(x,p) = \Psi[x,u(x,p)] \tag{10}$$

$$M(x,p) = -L[x,u(x,p)] + p'f(x,p) \tag{11}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix} \text{ is a constant vector and } u^* = u(x,p),$$

$u \in C^1$ is obtained by maximizing the Hamiltonian.

In this paper we solve the above TPBV problem by imbedding it in a class of TPBV problems. All members of the class are defined by (4-6) and differ only in the value of a parameter k introduced in the equation (3) of the terminal manifold τ,

$$\tau(k) = \left\{ x \in R^n,\ q(x,k) = \begin{bmatrix} q^1(x,k) \\ \vdots \\ q^m(x,k) \end{bmatrix} = 0 \right\}. \tag{12}$$

The manifold $\tau(k)$ is constructed so that the initial state x_o belongs to $\tau(k)$ at $k=0$ while at $k=1$, $\tau(k)$ is identical to the manifold τ of the original problem

$$x_o \epsilon \tau(0) \ , \quad \tau = \tau(1) \tag{13}$$

$$\text{rank } q_x(x,k) = m, \ x \epsilon \tau(k). \tag{14}$$

We start with the problem at $k=0$, which reduces to the solution of a system of algebraic equations, and we obtain the solution of the original TPBV problem at the end of a sequence of steps. Each step represents a transfer from the TPBV problem at a current value k to the neighboring TPBV problem at $k + \Delta k$.

To derive expressions for a step at k, we differentiate the equations of the TPBV problem

$$\dot{x}(t,k) = f[x(t,k),p(t,k)] \tag{15}$$

$$\dot{p}(t,k) = g[x(t,k),p(t,k)] \tag{16}$$

$$x(0,k) - x_o = 0 \tag{17}$$

$$p[T(k),k] - q'_x[x(T(k),k)] \cdot \mu(k) = 0 \tag{18}$$

$$q[x(T(k),k),k] = 0 \tag{19}$$

$$M[x_o,p(0,k)] = 0 \tag{20}$$

with respect to k and we obtain

$$\begin{bmatrix} \dot{\gamma}(t,k) \\ \dot{\delta}(t,k) \end{bmatrix} = \begin{bmatrix} f_x[x(t,k),p(t,k)] & f_p[x(t,k),p(t,k)] \\ g_x[x(t,k),p(t,k)] & g_p[x(t,k),p(t,k)] \end{bmatrix} \begin{bmatrix} \gamma(t,k) \\ \delta(t,k) \end{bmatrix} \tag{21}$$

$$\gamma(0,k) = 0 \tag{22}$$

$$\frac{dp_e(k)}{dk} - q'_x[x_e(k),k] \frac{d\mu}{dk} - \sum_{j=1}^{m} \mu_j(k) q^j_{xx}[x_e(k),k] \frac{dx_e(k)}{dk} - \frac{\partial q'_x[x_e(k),k]}{\partial k} \mu = 0 \tag{23}$$

$$q_x[x_e(k),k] \frac{dx_e(k)}{dk} + \frac{\partial q[x_e(k),k]}{\partial k} = 0 \tag{24}$$

$$\{p'(0,k)f_x[x_o,p(0,k)] - L_x[x_o,p(0,k)]\} \gamma(0,k) + \{p'(0,k)f_p[x_o,p(0,k)] - L_p[x_o,p(0,k)] + f'[x_o,p(0,k)]\} \delta(0,k) = 0 \tag{25}$$

where

$$\gamma(t,k) = \frac{\partial x(t,k)}{\partial k} \ , \quad \delta(t,k) = \frac{\partial p(t,k)}{\partial k} \ , \tag{26}$$

$$x_e(k) = x[T(k),k] \;, \quad p_e(k) = p[T(k),k]. \tag{27}$$

Using (21),

$$\begin{bmatrix} \dfrac{dx_e(k)}{dk} \\ \dfrac{dp_e(k)}{dk} \end{bmatrix} = \Phi[T(k),0,k] \begin{bmatrix} \gamma(0,k) \\ \delta(0,k) \end{bmatrix} + \frac{dT(k)}{dk} \begin{bmatrix} f[x_e(k),p_e(k)] \\ g[x_e(k),p_e(k)] \end{bmatrix} \tag{28}$$

where
$$\Phi[T(k),0,k] = \begin{bmatrix} \phi_{11}(k) & \phi_{12}(k) \\ \phi_{21}(k) & \phi_{22}(k) \end{bmatrix}$$

is the transition matrix of (21) evaluated at T(k). Note that in view of (22), only $\phi_{12}(k)$ and $\phi_{22}(k)$ are needed. We substitute (28) in (23-24) to obtain a system of linear equations in the remaining $n+m+1$ unknowns $\delta(0,k)$, $\frac{d\mu(k)}{dk}$ and $\frac{dT}{dk}$,

$$\{\phi_{22}(k) - \sum_{j=1}^{m} \mu_j(k) q^j_{xx}[x_e(k),k]\phi_{12}(k)\}\delta(0,k) - q'_x \frac{d\mu(k)}{dk} + \{g[x_e(k),p_e(k)] - \sum_{j=1}^{m} \mu_j(k) q^j_{xx}[x_e(k),k] f[x_e(k),p_e(k)]\} \frac{dT(k)}{dk} = \frac{\partial q'_x}{\partial k}\mu(k) \tag{29}$$

$$q_x[x_e(k),k]\phi_{12}(k)\delta(0,k) + q_x[x_e(k),k] f[x_e(k),p_e(k)] \frac{dT(k)}{dk} = -\frac{\partial q[x_e(k),k]}{\partial k} \tag{30}$$

$$\{p'(0,k) f_p[x_o,p(0,k)] + f'[x_o,p(0,k)] + L_p[x_o,p(0,k)]\}\delta(0,k) = 0 \tag{31}$$

Now we solve (18, 19, 20) at $k + \Delta k$. We use

$$p(0,k+\Delta k) \cong p(0,k) + \Delta k \cdot \delta(0,k) \tag{32}$$

$$\mu(k+\Delta k) \cong \mu(k) + \Delta k \frac{d\mu(k)}{dk} \tag{33}$$

$$T(k+\Delta k) \cong T(k) + \Delta k \frac{dT(k)}{dk} \tag{34}$$

as the initial guess and apply the modified Newton's method [5].

To initialize the process, $p(0,k)$, $\mu(k)$ and $T(k)$ are obtained by solving (13), (18), (20), and (30), at $k=0$, that is

$$T(0) = 0 \tag{35}$$

$$p(0,0) = q'_x[x(T(0)),0]\mu(0) \tag{36}$$

$$M[x_o, p(0,0)] = 0 \tag{37}$$

$$q_x[x_e(0),0]\phi_{12}(0)\delta(0,0) + q_x[x_e(0),0]f[x_e(0),p_e(0)]\,\frac{dT(0)}{dk} = -\,\frac{\partial q[x_e(0),0]}{\partial k}\,. \tag{38}$$

In view of (35), $x_e(0) = x_o$, $\phi_{12}(0) = 0$ and $p_e(0) = p(0,0)$. Replacing $p_e(0)$ in (37-38) by its value (36), a set of $m+1$ nonlinear algebraic equations is obtained,

$$q_x[x_o,0]f[x_o,q_x'(x_o,0)\mu(0)]\,\frac{dT(0)}{dk} = -\,\frac{\partial q(x_o,0)}{\partial k} \tag{39}$$

$$M[x_o,q_x'(x_o,0)\mu(0)] = 0 \tag{40}$$

which can be solved for $\mu(0)$ and $\frac{dT(0)}{dk}$. If $m > 1$, additional conditions are needed for the remaining unknowns $\delta(0,k)$ and $\frac{d\mu(k)}{dk}$ to be determined from (29) and (31) at $k=0$. They are obtained by differentiating (19) twice with respect to k,

$$\begin{aligned}
&2q_x^j[x_o,0]f_p[x_o,p(0,0)]\,\frac{dT(0)}{dk}\,\delta(0,0) + q_x^j[x_o,0]f[x_o,p(0,0)]\,\frac{d^2T(0)}{dk^2} \\
&\quad = -\left(\frac{dT(0)}{dk}\right)^2\Big\{f'[x_o,p(0,0)]q_{xx}^j(x_o,0)f[x_o,p(0,0)] \\
&\qquad + q_x^j(x_o,0)f_x[x_o,p(0,0)]f[x_o,p(0,0)] \\
&\qquad + q_x^j(x_o,0)f_p[x_o,p(0,0)]g[x_o,p(0,0)]\Big\} \\
&\quad - \frac{dT(0)}{dk}\left[\frac{\partial}{\partial k}\,q_x^j(x_o,0) + \left(\frac{\partial q^j(x_o,0)}{\partial k}\right)_x\right]f[x_o,p(0,0)] - \frac{\partial^2 q^j(x_o,0)}{\partial k^2} \\
&\qquad\qquad j = 1,\ldots,m.
\end{aligned} \tag{41}$$

The system (29), (31), and (41) can now be solved for $\delta(0,0)$, $\frac{d\mu(0)}{dk}$ and $\frac{d^2T(0)}{dk^2}$.

The manifold-imbedding algorithm is summarized by the flow chart in Figure 1. In the algorithm a prescribed maximum number N_{max} of modified Newton's steps (loop N in Fig. 1) is used to select Δk and improve the solution of (18-20) with the specified error, ϵ_{max}. Note that the coefficient matrix of (29-31) is the Jacobian to be used in the modified Newton's method.

It should be pointed out that the method is applicable if the following conditions are satisfied.

First, the integrand $L(x,u)$ of Eq. (2) should be nonnegative in $R^n \times R^s$ to guarantee (35). Second, the manifolds $\tau(k)$ must satisfy (14) at $k=0$ so that (39-40) can be solved for $\mu(0)$ and $\frac{dT(0)}{dk}$. Moreover $\frac{dT(0)}{dk} > 0$ is a useful necessary condition in case (36-37) has several solutions. Third, in view of (32-34) continuity of $\delta(0,k)$, $\frac{d\mu(k)}{dk}$ and $\frac{dT(k)}{dk}$ seems to be sufficient, however, at $k=0$, $\frac{d^2T(k)}{dk^2}$ must exist to initialize the method.

III. NUMERICAL RESULTS

The manifold-imbedding method was applied to the system

$$\dot{x}_1 = \cos x_3 \qquad x_1(0) = x_{01} \tag{42}$$

$$\dot{x}_2 = \sin x_3 \qquad x_2(0) = x_{02} \tag{43}$$

$$\dot{x}_3 = u \qquad x_3(0) = x_{03} \tag{44}$$

where u is the control and

$$J = \tfrac{1}{2}\int_0^T (1 + u^2)dt \tag{45}$$

is the performance index. The system (42-44) can represent the motion of an electron controlled by a transversal magnetic field [6], see Fig. 2.

Example 1. For this example, the terminal manifold is the unit circle

$$\tau = \{x \in R^3 : x_1^2 + x_2^2 - 1 = 0\} \tag{46}$$

for which a simple choice of $\tau(k)$ is the family of concentric circles

$$\tau(k) = \{x \in R^3 : x_1^2 + x_2^2 - 1 - (1-k)(x_{01}^2 + x_{02}^2 - 1) = 0\}. \tag{47}$$

From (35-38),

$$T(0) = 0 \tag{48}$$

$$p(0,0) = 2\mu_1(0)\begin{bmatrix} x_{01} \\ x_{02} \\ 0 \end{bmatrix} \tag{49}$$

$$\mu_1(0) = \frac{1}{4(x_{01}\cos x_{03} + x_{02}\sin x_{03})} \tag{50}$$

$$(x_{02}\sin x_{03} - x_{01}\cos x_{03})\frac{dT(0)}{dk} = -x_{01}^2 - x_{02}^2 + 1. \tag{51}$$

The equations (29-31) are

$$\left[\begin{array}{c|c|c}
 & -2x_{e1}(k) & -2\mu_1(k)\cos x_{e3}(k) \\ \cline{2-3}
\phi_{22}(k)-\mu_1(k)\begin{bmatrix}2 & & \\ & 2 & \\ & & 0\end{bmatrix}\phi_{12}(k) & -2x_{e2}(k) & -2\mu_1(k)\sin x_{e3}(k) \\ \cline{2-3}
 & 0 & \begin{array}{c} p_{e1}(k)\sin x_{e3}(k) \\ -p_{e2}(k)\cos x_{e3}(k)\end{array} \\ \hline
(2x_{e1}(k)\ 2x_{e2}(k)\ 0)\ \phi_{12}(k) & 0 & \begin{array}{c} 2x_{e1}(k)\cos x_{e3}(k)+ \\ 2x_{e2}(k)\sin x_{e3}(k)\end{array} \\ \hline
\cos x_{03}\ \sin x_{03}\ \ p_3(0,k) & 0 & 0
\end{array}\right]
\begin{bmatrix} \delta_1(0,k) \\ \delta_2(0,k) \\ \delta_3(0,k) \\ \dfrac{d\mu(k)}{dk} \\ \dfrac{dT(k)}{dk} \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 0 \\ -x_{01}^2-x_{02}^2+1 \\ 0 \end{bmatrix} \tag{52}$$

The 3×3 matrices $\phi_{12}(k)$ and $\phi_{22}(k)$ are obtained by the method of 'unit solutions' applied to (28). Initial states are chosen to satisfy $\frac{dT(0)}{dk} > 0$ in (50-51). Numerical results satisfying (18-20) within 10^{-8} are shown in Table 1. Resulting trajectories are given in Fig. 3. Steps (32-34) with values Δk listed in Table 1 were used along with modified Newton iterations. The average computation time was three minutes on a CDC 1604.

Example 2. In this example, the same system is considered but with a specified final time,

$$T = t_f, \tag{53}$$

in (45). By introducing a new state variable x_4 satisfying

$$\dot{x}_4 = 1 \qquad x_4(0) = 0 \tag{54}$$

and a new terminal manifold

$$\tau = \{x\varepsilon R^4,\ x_1^2 + x_2^2 - 1 = 0,\ x_4 - t_f = 0\} \tag{55}$$

the problem becomes time-free. A suitable choice for $\tau(k)$ is

$$\tau(k) = \{x\varepsilon R^4, x_1^2 + x_2^2 - 1 + (1-k)x_3 - (1-k)(x_{01}^2 + x_{02}^2 + x_{03} - 1) = 0,\ x_4 - kt_f = 0\}. \tag{56}$$

From (35-38)

Table 1

Case	x_o	$p(0,0)$	$\mu_1(0)$	$x_e(1)$	$p(0,1)$	$P_e(1)$	$T(1)$	$\mu_1(1)$	Δk
1	-1.	3.5156	-1.7578	.4720	-.3912	-.3912	2.056	-.4144	$\frac{1}{16}$
	2.	-7.0312		.8816	-.7307	-.7307			
	.4	0.		-1.139	-1.5132	-2×10^{-10}			
2	2.	-.5000	-.1250	.2487	-.1454	-.1454	2.085	-.2924	$\frac{1}{14}$
	2.	-.5000		.9686	-.5665	-.5665			
	3.1414	0.		3.9156	.8421	-10^{-8}			
3	-2.	.9254	-.2314	-.7588	.6476	.6476	1.438	-.4267	$\frac{1}{12}$
	0.	0.		.6513	-.5558	-.5558			
	1.	0.		.2355	-1.1116	3×10^{-12}			

$$\frac{dT(0)}{dk} = t_f \tag{57}$$

$$p_1(0,0) = \begin{bmatrix} 2\mu_1(0)x_{01} \\ 2\mu_1(0)x_{02} \\ \mu_1(0) \\ \mu_2(0) \end{bmatrix} \tag{58}$$

$$\mu_1(0) = \frac{-x_{01}^2 - x_{02}^2 + 1 - 2\,\frac{dT(0)}{dk}\,(x_{01}\cos x_{03} + x_{02}\sin x_{03})}{t_f}\,. \tag{59}$$

As shown by (61), $\mu_2(0)$ and $p_4(0,0)$ need not be computed. At $k=0$, eliminating $\delta_4(0,0)$, $\frac{d\mu_2(0)}{dk}$ and $\frac{dT(0)}{dk}$ from (29), (41), and (31), the system of equations reduces to:

$$\begin{bmatrix} 1 & 0 & 0 & -2x_{01} \\ 0 & 1 & 0 & -2x_{02} \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} \delta_1(0,0) \\ \delta_2(0,0) \\ \delta_3(0,0) \\ \frac{d\mu_1(0)}{dk} \end{bmatrix} = \begin{bmatrix} 2\mu_1(0)t_f\cos x_{03} \\ 2\mu_1(0)t_f\sin x_{03} \\ t_f[-p_1(0,0)\sin x_{03} + p_2(0,0)\cos x_{03}] - \mu_1(0) \\ -2t_f + 2\mu_1(0) \end{bmatrix} \tag{60}$$

After eliminating $p_4(0,k)$, $\mu_2(k)$, $\delta_4(0,k)$, $\frac{d\mu_2(k)}{dk}$ and $\frac{dT(k)}{dk} = t_f$ from (29-31), $\delta_j(0,k)$, $j = 1,\ldots,3$ and $\frac{d\mu_1(k)}{dk}$ are determined by

$$\begin{bmatrix} \phi_{22}(k) - \mu_1(k)\begin{bmatrix} 2 & & \\ & 2 & \\ & & 0 \end{bmatrix}\phi_{12}(k) & \begin{matrix} -2x_{e1}(k) \\ -2x_{e2}(k) \\ -1 + k \end{matrix} \\ (2x_{e1}(k) \;\; 2x_{e2}(k) \;\; 0)\; \phi_{12}(k) & 0 \end{bmatrix} \begin{bmatrix} \delta_1(0,k) \\ \delta_2(0,k) \\ \delta_3(0,k) \\ \dfrac{d\mu_1(k)}{dk} \end{bmatrix} =$$

$$\begin{bmatrix} 2\mu_1(k)t_f \cos x_{e3}(k) \\ 2\mu_1(k)t_f \sin x_{e3}(k) \\ t_f[p_{e2}(k)\cos x_{e3}(k) - p_{e1}(k)\sin x_{e3}(k) - \mu_1(k) \\ -(x_{01}^2 + x_{02}^2 + x_{03} - 1) - 2t_f[x_{e1}(k)\cos x_{e3}(k) + x_{e2}(k)\sin x_{e3}(k)] \\ + x_{e3}(k) - (1-k)t_f \end{bmatrix} \tag{61}$$

Numerical results are obtained for $k = 1$ for different initial states and final time, t_f. These are listed at $k = 1$ in Table 2. Typical trajectories are given in Fig. 4.

Table 2

Case	x_o	$p(0,0)$	$\mu_1(0)$	$x_e(1)$	$p(0,1)$	$p_e(1)$	$T(1)$	$\mu_1(1)$	Δk
1	-2.0 0.0 .0001	-13.60 0.0 3.40	3.400	-.8554 -.5431 -1.8461	-.8790 .5064 1.1353	-.8790 .5064 .1667	5.0	-.5265	.002
2	-2.0 0.0 .0001	-14.80 0.0 3.70	3.700	.4739 -.8816 -2.9973	-.1190 .2642 .6366	-.1190 .2642 .0873	10.0	-.1583	.001

IV. CONCLUSIONS

In the manifold-imbedding method a TPBV problem is reduced to a sequence of linear initial-value problems corresponding to a sequence of values of the imbedding parameter $k \in [0,1]$. These initial value problems are used to determine $\phi_{12}(k)$ and $\phi_{22}(k)$. Remaining operations needed for the solution of algebraic or transcendental equations require less computation time since Δk is kept within the domain of fast convergence of the modified Newton method.

ACKNOWLEDGMENTS

The authors wish to thank Professor Petar V. Kokotović for his guidance and helpful discussions in the preparation of the manuscript.

REFERENCES

1. Bryson, A. E. Jr. and Y. C. Ho, Applied Optimal Control, Blaisdell Publishing Co., Waltham, Mass., 1969.

2. Larson, R. E., "Survey of Dynamic Programming Computational Procedures," IEEE Trans. AC, Vol. AC-12, No. 6, 1967, pp. 767-774.

3. Kagiwada, K. H. and R. F. Kalaba, "Derivation and Validation of an Initial-Value Method for Certain Nonlinear Two-Point Boundary-Value Problems," JOTA, Vol. 2, No. 6, 1968, pp. 378-385.

4. Glasser, D., "Numerical Solution of Two-Point Boundary Value Problems on Total Differential Equations," SIAM J. Numer. Anal., Vol. 6, No. 4, December, 1969, pp. 591-597.

5. Kantorovich, L. V. and G. P. Akilov, Functional Analysis in Normed Spaces, Pergamon Press, New York, 1964.

6. Medanić, J. and C. I. Chen, "Feedback Strategies for Differential Game with Singular Solutions," Proc. 1st Int. Conf. on the Theory and Applications of Diff. Games, October, 1969, pp. VII 8-13.

FIGURES

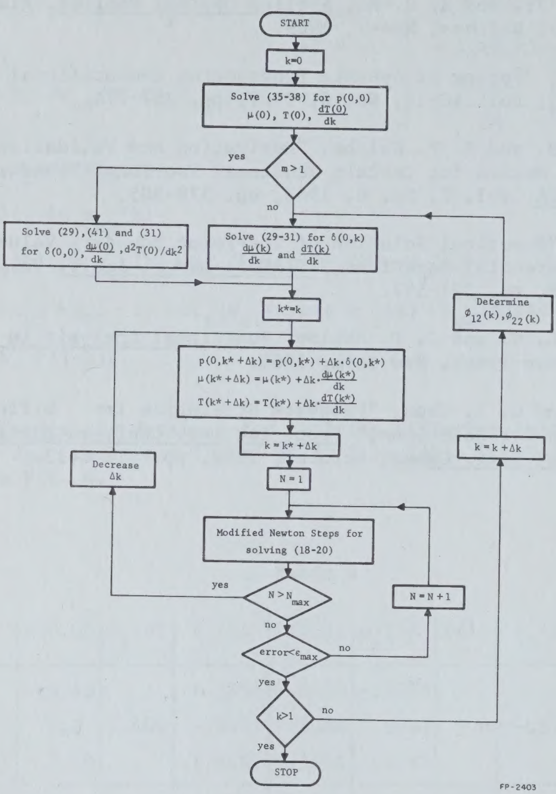

START
k=0
Solve (35-38) for p(0,0)
$\mu(0)$, T(0), $\frac{dT(0)}{dk}$
m > 1
yes
no
Solve (29),(41) and (31)
for $\delta(0,0)$, $\frac{d\mu(0)}{dk}$, $d^2T(0)/dk^2$
Solve (29-31) for $\delta(0,k)$
$\frac{d\mu(k)}{dk}$ and $\frac{dT(k)}{dk}$
k*=k
Determine
$\phi_{12}(k), \phi_{22}(k)$
$p(0,k^*+\Delta k) = p(0,k^*) + \Delta k \cdot \delta(0,k^*)$
$\mu(k^*+\Delta k) = \mu(k^*) + \Delta k \cdot \frac{d\mu(k^*)}{dk}$
$T(k^*+\Delta k) = T(k^*) + \Delta k \cdot \frac{dT(k^*)}{dk}$
k = k* + Δk
k = k + Δk
Decrease
Δk
N = 1
Modified Newton Steps for
solving (18-20)
yes
N > N_{max}
N = N + 1
no
error<ϵ_{max}
no
yes
k>1
no
yes
STOP
FP-2403

Figure 1. A flow chart for the manifold-imbedding method.

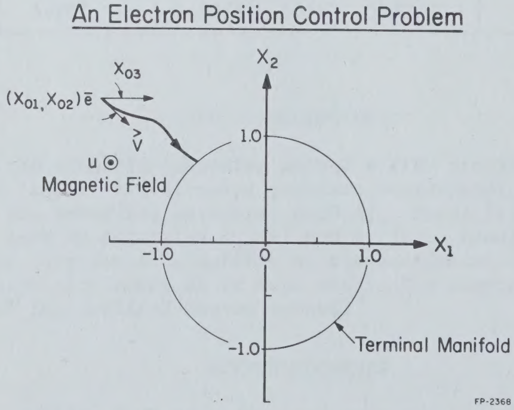

An Electron Position Control Problem
X_2
X_{03}
$(X_{01}, X_{02})\bar{e}$
$\vec{V}$
u ⊙
Magnetic Field
1.0
-1.0
0
1.0
X_1
-1.0
Terminal Manifold
FP-2368

Figure 2. (FP-2368) An electron optimal position control problem.

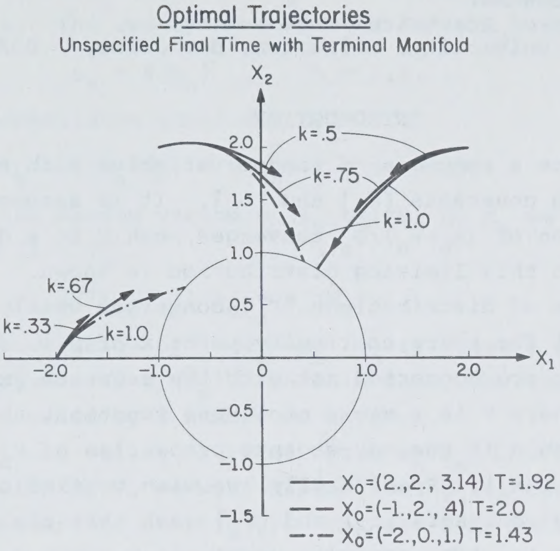


Figure 3. (FP-2364) Optimal trajectories of Example 1.

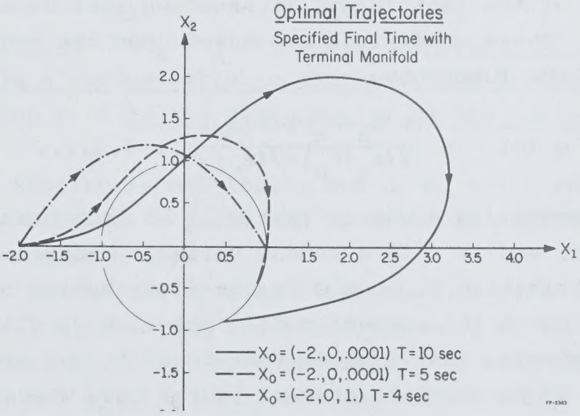


Figure 4. (FP-2363) Optimal trajectories of Example 2.

NONLINEAR EXTENSIONS OF A LIMIT THEOREM

ROBERT LUGANNANI
Department of Electrical Engineering
Princeton University, Princeton, New Jersey 08540

INTRODUCTION

Let $\{x_n\}$ be a sequence of random variables with centering and normalizing constants $\{a_n\}$ and $\{b_n\}$. It is assumed that the distribution of $(x_n-a_n)/b_n$ converges weakly to a distribution F and that this limiting distribution is known. (Recall that a sequence of distributions $\{F_n\}$ converges weakly to F if $\lim_n F_n(x)=F(x)$ for every continuity point x of F.) In many applications we are concerned not with the sequence $\{x_n\}$ but with $\{V(x_n)\}$ where V is a given nonlinear function, and it would be desirable if the convergence properties of x_n could be extended to $V(x_n)$. Specifically, we wish to find centering and normalizing constants $\{\alpha_n\}$ and $\{\beta_n\}$ such that the distributions of $(V(x_n)-\alpha_n)/\beta_n$ converge weakly to a limit G and to evaluate this limit. The solution of this problem is useful in estimating the distribution of $V(x_n)$ for large but finite values of n.

In the present paper, sufficient conditions are given for the existence of the limit G and it is shown that G is simply related to F. These conditions are based upon the convergence properties of the functions

$$g_n(y) = \frac{V(a_n+b_ny)-V(a_n)}{V(a_n+b_n)-V(a_n)}, \quad n=1,2,\ldots . \qquad (1)$$

If $g_n(y)$ converges (in a manner that will be made precise) to a function $g(y)$ and if x is a random variable characterized by the limit distribution F, then G is the distribution of $g(x)$. For many nonlinearities encountered in practice we will find, rather surprisingly, that $g(y)$ is restricted to one of two basic forms. Of particular interest is the case where $g(y)=y$, which implies that F and G are identical. The variety of limit laws that can arise are illustrated by several examples.

THE EXISTENCE OF G AND ITS RELATIONSHIP TO F

We require throughout that V not be constant. This requirement eliminates from consideration the uninteresting case where the distribution of $V(x_n)$ is degenerate and identical for each n. The centering constants for $V(x_n)$ are given by

$$\alpha_n = V(a_n) \quad , \qquad n = 1,2,... \tag{2}$$

and the normalizing constants by

$$\beta_n = V(a_n+b_n) - V(a_n) \quad , \qquad n = 1,2,... \quad . \tag{3}$$

Denoting the random variable $(x_n-a_n)/b_n$ by $\hat{x}_n$ we have, upon combining (1)-(3),

$$(V(x_n)-\alpha_n)/\beta_n = g_n(\hat{x}_n) \quad ,$$

and the convergence of $(V(x_n)-\alpha_n)/\beta_n$ is clearly equivalent to the convergence of $g_n(\hat{x}_n)$. In this form, the limit problem has been studied by a number of authors.

For those values of y for which a_n+b_ny lies in the domain of definition of V, we will assume that $g_n(y)$ converges to a (Borel) measurable limit $g(y)$. The set E is defined as the collection of points, y, for which $g_n(y_n)$ does not converge to $g(y)$ for some sequence $y_n \to y$. When E is a null set with respect to the probability measure induced by the distribution F, we shall write $P_F(E)=0$. The basic convergence properties for $(V(x_n)-\alpha_n)/\beta_n$ are given by

THEOREM 1. Let $g_n(y) \to g(y)$ and let $P_F(E)=0$. Then the distribution of $(V(x_n)-\alpha_n)/\beta_n$ converges weakly to a limit G, and if x is a random variable characterized by the limiting distribution F, G is the distribution of the random variable $g(x)$.

This theorem is originally due to H. Rubin whose results remain unpublished; a proof can be found in Billingsley [1, p. 34] or Anderson [2].

Let D be the set of discontinuities of $g(y)$. If $g_n(y) = g(y)$ for all n, then E=D and the hypothesis of the theorem requires that $P_F(D)=0$. More generally, we will have $P_F(E)=0$ if $P_F(D)=0$ and $g_n(y)$ converges uniformly on every closed bounded interval. These conditions frequently proove useful in practice; unfortunately , they are stated in terms of the auxiliary functions $g_n(y)$ and our primary concern is with the nonlinearity V. In the remainder of this section we shall examine conditions on V which ensure that Theorem 1 holds.

Suppose that x_n is the sum of n independent random variables each with mean μ and standard deviation σ. If we take $a_n=n\mu$ and $b_n=\sqrt{n}\,\sigma$, it is well known that the distribution of $(x_n-a_n)/b_n$ converges to a normal law with zero mean and unit variance. Motivated by this elementary version of the central limit theorem, we shall consider two cases. The first case corresponds to a non-zero value for μ and we have $a_n\uparrow\infty$ or $a_n\downarrow-\infty$ and $(b_n/a_n)\to 0$, while the second case corresponds to $\mu=0$ for which $a_n=0$ and $b_n\uparrow\infty$. In addition to the above examples these two situations are important in a variety of applications. For both cases we will determine the limiting function g(y) and the set E needed to apply Theorem 1. Under rather mild restrictions on the nonlinearity V, we will find that g(y) has a very simple form and that E is easily described

For the first case we have

THEOREM 2. Let $a_n\uparrow\infty$ $(a_n\downarrow-\infty)$ and let $(b_n/a_n)\to 0$. If V has a derivative, V', and if

$$\lim_{\lambda\uparrow\infty}\frac{V'(\lambda y)}{V'(\lambda)}\qquad\left(\lim_{\lambda\downarrow-\infty}\frac{V'(\lambda y)}{V'(\lambda)}\right)\tag{4}$$

exists and is measurable for all $y>0$, then

$$g_n(y)\to g(y)=y\tag{5}$$

and the set E is empty.

Condition (4) is simply a statement that V' is a function of regular variation. These functions were introduced by Karamata [3] and subsequently studied by many authors; perhaps the most accessible treatment of their properties can be found in Feller [4]. While it is usually required that functions of regular variation be nonnegative and defined on $(0,\infty)$, it is a simple matter to extend the definition to functions that take on negative values and are defined on $(-\infty,0)$. For our purposes we need only one result from the general theory of functions of regular variation, namely, that when either of the limits (4) exist and are measurable, the convergence is uniform on every bounded interval not containing the origin and the limit is given by x^γ where γ is a real constant.

Proof of Theorem 2. Using the mean value theorem we may write

$$g_n(y) = \int_{a_n}^{a_n+b_ny} V'(s)ds \Big/ \int_{a_n}^{a_n+b_n} V'(s)ds$$

$$= y \cdot V'(\eta_n \delta_n(y)) \,/\, V'(\eta_n) \tag{6}$$

where $a_n \le \eta_n \le a_n + b_n$ and

$$\frac{1}{1+(b_n/a_n)} \le \delta_n(y) \le 1 + (b_n/a_n)y \quad .$$

Depending upon whether $a_n \uparrow \infty$ or $a_n \downarrow -\infty$ we have $\eta_n \uparrow \infty$ or $\eta_n \downarrow -\infty$ and $\delta_n(y) \to 1$. Combining these properties with the uniform convergence of (4) we obtain

$$\lim_n V'(\eta_n \delta_n(y)) \,/\, V'(\eta_n) = 1$$

and consequently $g_n(y) \to g(y) = y$.

To show that the set E is empty, let $y_n \to y$ and observe that

$$|g_n(y_n) - g(y)| \le |g_n(y) - g(y)| + |g_n(y_n) - g_n(y)| . \tag{7}$$

The first term on the right side of this inequality goes to zero as $n \to \infty$ because of the convergence of $g_n(y)$. For the second term we again use the mean value theorem to establish

$$|g_n(y_n) - g_n(y)| = |y_n - y| \cdot |V'(\eta_n \epsilon_n(y, y_n)) \,/\, V'(\eta_n)|$$

where

$$\frac{1+(b_n/a_n)y_n}{1+(b_n/a_n)} \le \epsilon_n(y, y_n) \le 1 + (b_n/a_n)y$$

and η_n is the same as before. Clearly $\epsilon_n(y, y_n) \to 1$, and

$$\lim_n |V'(\eta_n \epsilon_n(y, y_n)) \,/\, V'(\eta_n)| = 1$$

as a consequence of the uniform convergence.

Now because $y_n \to y$, it follows that the second term on the right side of (7) goes to zero as $n \to \infty$. Thus, $g_n(y_n) \to g(y)$, which implies that E is empty and the proof is complete.

The implication of Theorem 2 is that whenever V' is a function of regular variation, the limiting distributions F and G are identical. It is, of course, not necessary that V' satisfy (4) for this conclusion to hold, since $g_n(y)$ may converge to y for some sequences $\{a_n\}$ and $\{b_n\}$ but not for others. Nevertheless many nonlinearities encountered in practice do have derivatives that satisfy (4) and in these cases the theorem provides an extremely simple characterization of the limit G.

Perhaps the simplest class of functions that satisfy (4) are those whose derivative, $V'(x)$, approaches a constant as $x\uparrow\infty$ or $x\downarrow-\infty$. We then have, for $y>0$,

$$\lim_{\lambda\uparrow\infty}\frac{V'(\lambda y)}{V'(\lambda)} = \lim_{\lambda\downarrow-\infty}\frac{V'(\lambda y)}{V'(\lambda)} = 1\ . \tag{8}$$

Rational functions form another important class. If

$$V(x) = \sum_{n=0}^{N} c_n x^n \bigg/ \sum_{m=0}^{M} d_m x^m \tag{9}$$

it is easily seen that

$$\lim_{\lambda\uparrow\infty}\frac{V'(\lambda y)}{V'(\lambda)} = \lim_{\lambda\downarrow-\infty}\frac{V'(\lambda y)}{V'(\lambda)} = \begin{cases} y^{N-M-1}, & N\neq M \\ y^{L-N-1}, & N=M \end{cases} \tag{10}$$

where $0\le L\le N-1$ is the largest integer that satisfies $c_N d_L \neq d_N c_L$ (the existence of such an integer is guaranteed by the assumption that V is not equal to a constant). As an example of a function that is not rational consider

$$V(x) = \frac{x}{1+|x|} \tag{11}$$

In this case

$$\lim_{\lambda\uparrow\infty}\frac{V'(\lambda y)}{V'(\lambda)} = \lim_{\lambda\downarrow-\infty}\frac{V'(\lambda y)}{V'(\lambda)} = y^{-2}\ . \tag{12}$$

This particular function will be used in the examples of the next section.

Next we turn to the case where $a_n=0$ and $b_n\uparrow\infty$. Define $\mathrm{sgn}[x]$ to be +1,0, or −1 according as $x>0$, $x=0$, or $x<0$, and define $K(\pm 1)$ by $K(1)=1$ and

$$K(-1) = \lim_{\lambda\uparrow\infty}\frac{V(-\lambda)-V(0)}{V(\lambda)-V(0)}$$

provided this limit exists. The pertinent results for this case are contained in

THEOREM 3. *Let* $a_n=0$ *and* $b_n\uparrow\infty$. *If, for* $y\neq 0$, *the limit*

$$\lim_{\lambda\uparrow\infty} \frac{V(\lambda y) - V(0)}{V(\lambda) - V(0)} \tag{13}$$

exists and is measurable, then for some real γ

$$g_n(y)\to g(y) = \begin{cases} K(\operatorname{sgn}[y])|y|^{\gamma}, & y\neq 0 \\ 0, & y=0 \end{cases} \quad . \tag{14}$$

When $\gamma\leq 0$, *the set* E *consists of the single point* $\{0\}$, *while* if $\gamma>0$, E *is empty*.

Again we are concerned with functions of regular variation; this time, however, it is the function itself and not its derivative that is of interest. If $\hat{V}(x)=V(x)-V(0)$, the hypothesis of the theorem states that $\hat{V}(\lambda y)/\hat{V}(\lambda)$ converges. When $y>0$, we have already noted that this convergence is uniform and the limit is given by y^{γ}. When $y<0$, write

$$\frac{\hat{V}(\lambda y)}{\hat{V}(\lambda)} = \frac{\hat{V}(-\lambda|y|)}{\hat{V}(\lambda|y|)} \cdot \frac{\hat{V}(\lambda|y|)}{\hat{V}(\lambda)} \to K(-1)|y|^{\gamma}$$

where again the convergence is uniform for $y\neq 0$.

Proof of Theorem 3. From the above remarks it is clear that (14) holds for $y\neq 0$, while for $y=0$, the desired result is a consequence of the fact that $\hat{V}(0)=0$. Since $g_n(y)$ converges uniformly for $y\neq 0$ and $g(y)$ is continuous, it follows that no point $y\neq 0$ can belong to E. To establish the conditions under which $\{0\}\in E$, a lengthy analysis is required involving the canonical representation of functions of regular variation. This aspect of the proof is beyond the scope of the present paper and will be omitted.

For those situations covered by Theorem 3, the limiting distribution G is again very simply related to F, this relationship being governed by a power law with exponent γ on each of the semi-infinite intervals $(-\infty,0)$ and $(0,\infty)$. As in the previous case the existence of the limit (13) is sufficient but not necessary.

If $V(x)$ converges to a finite limit as $x\uparrow\infty$ or $x\downarrow -\infty$, the desired limit exists and $\gamma=0$. For the rational function (9) we find that when $y\neq 0$ and $c_o\neq 0$

$$\lim_{\lambda\uparrow\infty} \frac{V(\lambda y)-V(0)}{V(\lambda)-V(0)} = \begin{cases} y^{N-M}, & N>M \\ y^{L-N}, & N=M \\ 1, & N<M \end{cases} \tag{15}$$

where $1 \leq L \leq N$ is the largest integer satisfying $c_o d_L \neq d_o c_L$ (such an integer exists because $V(x)$ is not constant). When $y \neq 0$ and $c_o = 0$

$$\lim_{\lambda \uparrow \infty} \frac{V(\lambda y) - V(0)}{V(\lambda) - V(0)} = y^{N-M} \quad . \tag{16}$$

In both cases the limits may be written in the form (14) by defining $K(-1)$ and γ in an obvious manner. Finally, for the function (10) we have

$$\lim_{\lambda \uparrow \infty} \frac{V(\lambda y) - V(0)}{V(\lambda) - V(0)} = \operatorname{sgn}[y], \tag{17}$$

and this will have the desired form if we choose $\gamma = 0$ and $K(-1) = -1$.

APPLICATIONS

The foregoing results will be applied to four examples that illustrate the variety of limiting distributions that can arise. In each example the sequence $\{x_n\}$ consists of a simple modification of the sum $\Sigma_i^n \xi_i^2$ where the ξ_i's are independent, Gaussian variables with zero mean and unit variance. This sum is a chi-squared random variable with n degrees of freedom; its mean and standard deviation are n and $\sqrt{2n}$ respectively. The nonlinearity V will have the form

$$V(x) = \frac{x}{1 + |x|} \quad .$$

Both $V(x)$ and $V'(x)$ are functions of regular variation whose respective limits are given by (17) and (12).

By selecting different centering and normalizing constants for x_n, the limiting distribution F and the limit $g(y)$ can be altered to provide a variety of limiting distributions G. The centering and normalizing constants for $V(x_n)$ are given by (2) and (3), and the distribution of $(V(x_n) - \alpha_n)/\beta_n$ will be denoted by G_n.

Example 1. For this example x_n is given by

$$x_n = \sum_{i=1}^{n} \xi_i^2 \quad .$$

If we choose $a_n = n$ and $b_n = \sqrt{2n}$, the distribution of $(x_n - a_n)/b_n$ is known to converge to a normal law with zero mean and unit variance. Since $a_n \uparrow \infty$ and $(b_n/a_n) \to 0$, the pertinent results are given by Theorem 2 and we see that G_n also has a normal limit with zero mean and unit variance. The distribution G_n together with its limiting value is presented in Fig. 1.

Example 2. The random variable x_n is again given by

$$x_n = \sum_{i=1}^{n} \xi_i^2 \quad .$$

However in this case we choose $a_n=0$ and $b_n=n$, which implies that the limiting distribution of $(x_n-a_n)/b_n$ will be degenerate with its point of increase at +1. Actually, an even stronger result holds, viz., that $(x_n-a_n)/b_n \to 1$ with probability 1. Theorem 3 applies to this example with $g(y)=\mathrm{sgn}[y]$, and, since $P_F(E)=P_F(\{0\})=0$, G_n also has a degenerate limit with its point of increase +1. The associated curves are given in Fig. 2.

Example 3. Here we consider the zero-mean random variable

$$x_n = \sum_{i=1}^{n} \xi_i^2 - n \quad .$$

By selecting $a_n=0$ and $b_n=\sqrt{2n}$, the limiting distribution of $(x_n-a_n)/b_n$ is seen to be normal with zero mean and unit variance. As in Example 2, $g(y)=\mathrm{sgn}[y]$ and $P_F(E)=P_F(\{0\})=0$. This implies (Theorem 3) that G_n has a limit which is constant except for jumps of 1/2 at ±1. The pertinent details are displayed in Fig. 3.

Example 4. This last example is concerned with the modified sum

$$x_n = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} \xi_i^2 \quad .$$

If $a_n=\sqrt{n/2}$ and $b_n=1$, the distribution of $(x_n-a_n)/b_n$ converges to a normal law with zero mean and unit variance. This example is covered by Theorem 2 and the limiting distribution for G_n is normal with zero mean and unit variance. Figure 4 contains the related curves.

ACKNOWLEDGEMENT

This work was supported in part by the USAF Office of Scientific Research under AFOSR Grant 1333-67, in part by the National Science Foundation under Grant GK-1439, and in part by the U.S. Army Research Office under Contract DAHC04-69-0012.

REFERENCES

[1] P. Billingsley, Convergence of Probability Measures, Wiley, New York, 1968.

[2] T.W. Anderson, "Asymptotic Theory for Principal Component Analysis," Annals Math. Statist., 34(1963), pp. 122-148.

[3] J. Karamata, "Sur un Mode de Croissance Regulière des Fonctions," Mathematica (Cluj), 4 (1930), pp. 38-53.

[4] W. Feller, An Introduction to Probability Theory and Its Applications (Vol. II), Wiley, New York, 1966.

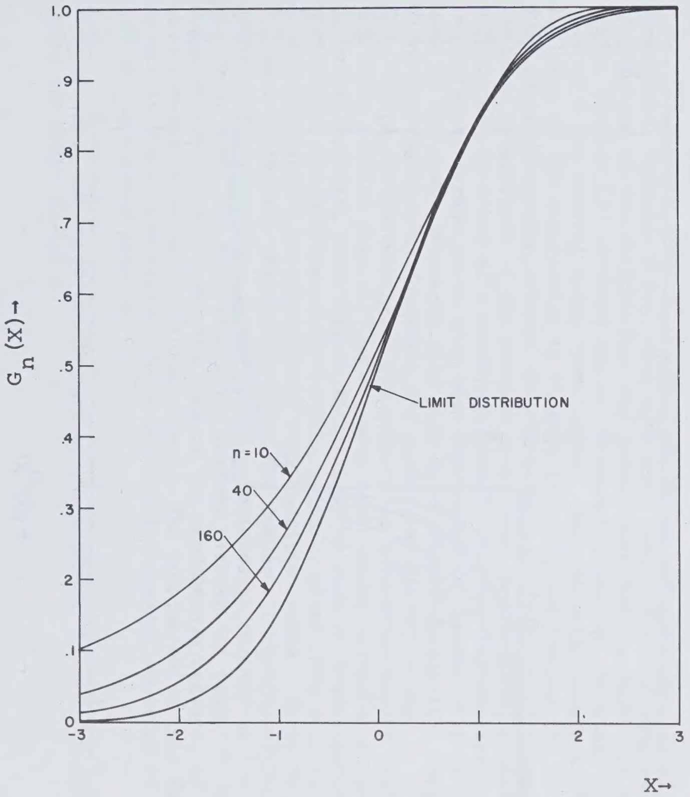


Fig. 1 Example of convergence to limit law: $g_n(y) \rightarrow y$.

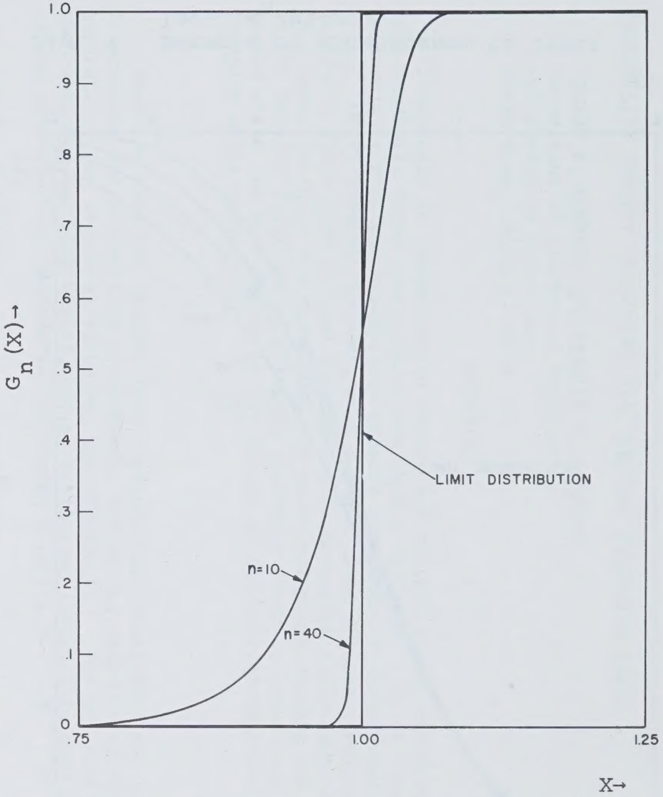


Fig. 2 Example of convergence to limit law: $g_n(y) \rightarrow \mathrm{sgn}[y]$.

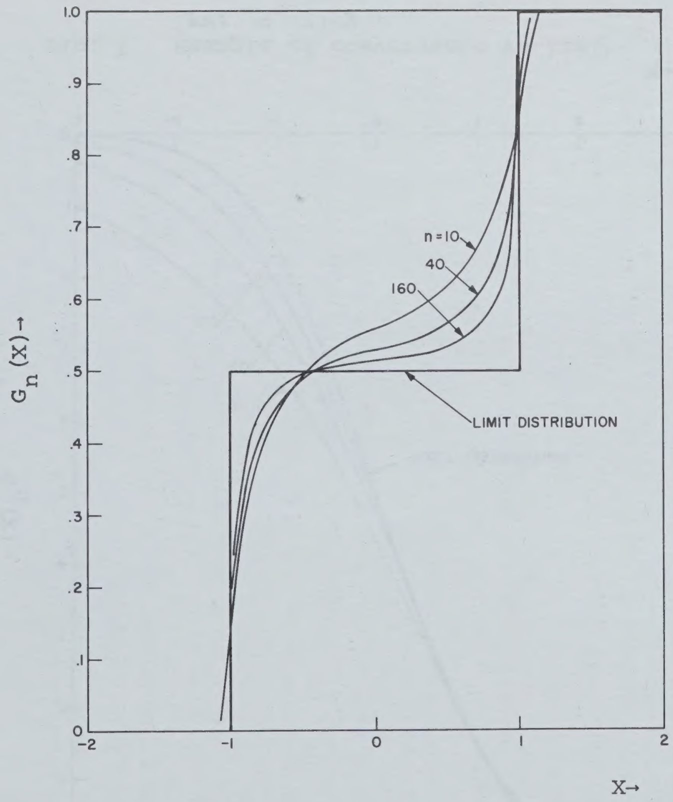


Fig. 3   Example of convergence to limit law: $g_n(y) \rightarrow \text{sgn}[y]$.

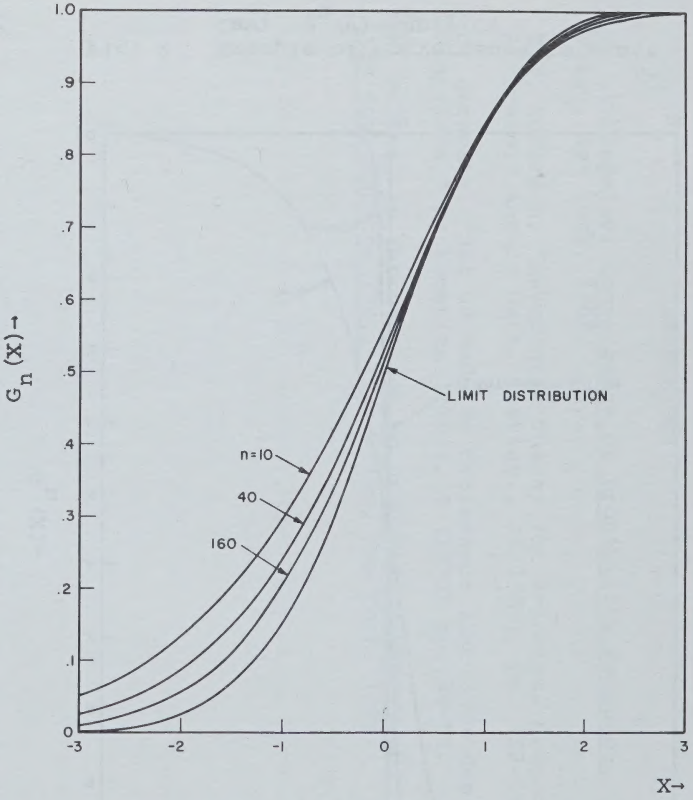


Fig. 4   Example of convergence to limit law: $g_n(y) \rightarrow y$.

PROBABILITY DENSITY TECHNIQUES FOR INTEGRAL QUADRATIC FORMS

JOSEPH K. MURDOCK AND STANLEY R. LIBERTY
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana 46556

ABSTRACT

Frequently in communication theory and control theory contexts integral quadratic forms appear which are random variables. It is often desirable to know the probability density functions of these random variables. Until present, however, only a small class of these integral quadratic forms have yielded to attempts to determine their densities. This paper contains two techniques for numerically determining the probability density functions of integral quadratic forms in Gauss-Markov processes which can be modelled by the output of linear systems driven by Gaussian white noise.

Over the past two decades investigators have considered the problem of determining the probability distribution and/or the probability density of quadratic forms in normal variates; see example [1]. The motivation for such research has been varied, but of particular interest in many communication and detection systems is the random quantity,

$$Z = \frac{1}{T} \int_{-T/2}^{T/2} x^2(t)dt$$

i.e., the average power of a sample of finite time duration, T, of a one-dimensional Gaussian noise, x(t) with zero mean. Rice [2] obtained an approximation to the distribution and expressions for the first few moments of Z. Slepian [3] obtained series representations for the exact distribution of Z for several noise covariances. Very recently Schwartz [4], building on the observation of Siegert [5], [6] and Slepian [3] that there is a close connection between the characteristic function of Z and the Fredholm determinant of the process covariance, has obtained a closed-form expression, in terms of a resolvant kernel, for the characteristic function of the time-average power of a zero mean Gaussian process that is continuous in the mean. Schwartz also gave examples with analytical approximations for the density functions and cumulative distributions.

To the best knowledge of the authors no literature exists dealing with the above problem when x(t) is non-zero mean and/or multidimensional. There is, however, considerable motivation for such an investigation, in particular, the problem of controlling a linear plant with a linear feedback controller in the presence of state additive noise. There is, normally, associated with this problem, for the purposes of optimization, some measure of performance. A very popular performance measure is the integral quadratic cost. In this stochastic control problem the state of the system, which appears in the quadratic form of the performance measure, is in general non-zero mean and multidimensional. The problem which this paper treats is the determination of the probability density function of such a random performance measure.

This research was supported in part by the National Science Foundation under the Grant GK-13618.

REFERENCES

1. U. Grenander, H.O. Pollak, and D. Slepian, "The Distribution of Quadratic Forms in Normal Variates: A Small Sample Theory with Applications to Spectral Analysis," J. SIAM, Vol. 7, December 1959.
2. S.O. Rice, "Mathematical Analysis of Random Noise," Bell Sys. Tech. J., Vol. 23 and 24, Sec. 3.9; also in Noise and Stochastic Processes. New York: Dover, 1954.
3. D. Slepian, "Fluctuations of Random Noise Power," Bell Sys. Tech. J., January 1958.
4. M.I. Schwartz, "Distribution of the Time-Average Power of a Gaussian Process," IEEE Trans. Info Theory, Vol.IT-16, pp.17-26, Jan. 1970.
5. A.J.F. Siegert, "Passage of Stationary Processes Through Linear and Nonlinear Devices," IRE Trans. Info. Theory, Vol. IT-3, pp. 4-25, March 1954.
6. ______, "A Systematic Approach to a Class of Problems in the Theory of Noise and Other Random Phenomena - Part II, Examples," IRE Trans. Info. Theory, Vol. IT-3, pp. 38-43, March 1957.

ON THE APPLICATIONS OF RANDOM SIGNALS WITH MULTIPLICITY ONE

ANTHONY EPHREMIDES and JOHN B. THOMAS
Department of Electrical Engineering
Princeton University, Princeton, New Jersey 08540

ABSTRACT

In a number of applications of random signal theory, it is necessary to consider random processes with multiplicity one. Such processes possess the property that they may be "prewhitened" by a linear, causal, invertible filter. After presenting some background on the notion of multiplicity, this paper considers two important applications: (1) The modeling of stochastic signals by linear systems, and (2) the prewhitening approach to linear least-mean-square estimation.

I. MATHEMATICAL BACKGROUND

The class C of random processes considered in this paper consists of all zero mean, second order processes $X(t)$ that are left mean square continuous. The closed linear manifold spanned by $X(s)$, $s \leq t$, is a separable Hilbert space denoted by $H(X,t)$. The space $H(X,+\infty)$ is denoted simply by $H(X)$, while $H(X,-\infty)$ denotes $\bigcap_t H(X,t)$. Cramer [1,2] and Hida [3] have shown that, if $X(t) \in C$, there exists a unique orthogonal decomposition of the form

$$X(t) = U(t) + V(t) \quad , \quad \text{with } H(U) \perp H(V)$$

where $U(t)$ is deterministic, that is $H(U,-\infty)=H(U,t)$, $\forall t \in R$, and $V(t)$ is purely non-deterministic (pnd), that is $H(V,-\infty)=0$. Furthermore, the pnd portion $V(t)$ has a canonical decomposition of the form

$$V(t) = \sum_{i=1}^{M} \int_{-\infty}^{t} g_i(t,\tau)\, dz_i(\tau) \quad , \quad \forall t \in R$$

where the $\{z_i(\tau)\}$ are mutually orthogonal processes in C with orthogonal increments, with $E[dz_i(\tau)]^2 = dF_i(\tau)$, and such that

$$H(V,t) = H(z_1,t) \oplus \dots \oplus H(z_M,t) \quad , \quad \forall t \in R$$

It is possible to choose the $\{z_i(\tau)\}$ so that the measures F_i are finite and form a chain of absolute continuity (i.e., $F_1 >> \dots >> F_M$). The number M, which may be any non-negative integer or ∞, is uniquely determined by the autocovariance of $X(t)$ and is called the multiplicity of the process.

As Mandrekar and Kallianpur have pointed out [4,5], the multiplicity of such a random process coincides with the spectral multiplicity of an operator A acting on $H(X)$ and whose resolution of the identity is P_t [the family of projection operators acting on $H(X)$ and with range $H(X,t)$]. In other words M is the dimension of the minimal generating subspace of A [6].

II. PRELIMINARY REMARKS

In problems of detection, estimation and optimization, it is often useful to model a random process $X(t)\in C$ as the output of a causal linear filter with white noise input $n(t)$. It is often necessary that this representation be invertible; that is, that the white noise can be recovered by passing $X(t)$ through another causal, linear filter. It is customary to express the process $X(t)$ by the formula

$$X(t) = \int_{-\infty}^{t} h(t,\tau)\, n(\tau)\, d\tau \tag{1}$$

Since there are some difficulties in interpreting and using this integral we prefer the following equivalent notation

$$X(t) = \int_{-\infty}^{t} h(t,\tau)\, dz(\tau) \tag{2}$$

where $z(\tau)$ is a process with orthogonal increments. The integral in Eq. (2) is an ordinary stochastic integral [7]. It is always possible to choose $z(\tau)$ to be a pnd process. If by white noise is meant a Gaussian, wide-sense stationary process with constant spectrum, then $z(\tau)$ is Brownian motion. The invertibility requirement is then expressed by $H(X,t) = H(z,t)$, $\forall t\in R$. To avoid unnecessary complications we may choose $z(\tau)$ to be also a second-order process in which case the corresponding real measure F defined by $dF(\tau) = E[dz(\tau)]^2$ is finite. It is apparent that, in order to specify conditions under which $X(t)$ can be modeled in this way, it is necessary and sufficient to determine conditions for $X(t)$ to be a pnd random process with multiplicity $M=1$. In a different terminology we say that $X(t)$ is linearly equivalent to white noise.

In linear, least-mean-square estimation, when the observation process $y(t)$ is the sum of a random signal $z(t)$ and uncorrelated noise $n(t)$, it is convenient to use a prewhitening technique [8] whereby $y(t)$ is passed through a linear filter that produces a white noise output $\nu(t)$ which is then passed through a linear estimator. This method simplifies the resulting mathematical expression and provides some insight into the mechanism of estimation. To avoid the difficulties arising from the use of $n(t)$ when the noise is white we shall consider the equivalent problem of estimating $Z(t)$ from $Y(s) = Z(s)+N(s)$, $s\leq t$ where y, z and n are the formal derivatives of Y, Z, and N. The noise process $N(t)$ in this case is a process with orthogonal increments and can be chosen to be pnd and second-order. It is of interest to specify the class of processes for which this approach is valid, particularly in view of recent extensions of the method [9,10]. Since the prewhitening must be carried out without loss of information and since only causal filters are usually considered, the problem is again equivalent to the requirement that $Y(t)$ be a pnd random process with multiplicity $M=1$.

The remainder of this paper is divided into two major sections each dealing with one of these two problems.

III. LINEAR EQUIVALENCE TO WHITE NOISE

Part I: Conditions on the Impulse Response Function

Let $X(t)\in C$ be the output of a causal, linear filter with a white noise input, that is, $X(t) = \int_{-\infty}^{t} h(t,\tau)dz(\tau)$. The question in this case is whether $H(X,t) = H(z,t)$, $\forall t\in R$. Levy [11] has shown that if $h(t,t)$ is continuous and not zero in an interval $[a,b]$ and if $\partial h(t,\tau)/\partial\tau$ exists and is bounded in that interval, then $H(X,t) = H(z,t)$, $\forall t\in[a,b)$ where $z(t)$ is Brownian motion. The following theorem generalizes this result.

Theorem 1: Let $h(t,\tau)$ be the impulse response function of a causal, linear system. If $h(t,t)$ is continuous and not zero in $[a,b]$ and if $\partial h(t,\tau)/\partial\tau$ exists and is continuous in that interval, then $H(X,t) = H(z,t)$, $\forall t\in[a,b)$ for any input process $z(t)\in C$ with orthogonal increments.

Proof: For $t \in [a,b]$, we have

$$X(t) = \int_a^t h(t,\tau)\, dz(\tau) \quad , \quad E[dz(\tau)]^2 = dF(\tau)$$

It is true that $H(X,t) \subset H(z,t)$, since $h(t,\tau)$ is not a σ-function [11]. It will suffice therefore to show that $u\in H(z,t)$ and $u \perp H(X,t)$ imply $u=0$. We have

$$u \in H(z,t) \Rightarrow u = \int_a^t g(\tau)\, dz(\tau) \text{ for some } g(\tau)\in L_2(dF;a,t)$$

$$u \perp H(X,t) \Rightarrow E[u\cdot X(s)] = 0 \quad , \quad \forall\, s \in[a,t]$$

that is

$$\int_a^s h(s,\tau)\, g(\tau)\, dF(\tau) = 0 \quad , \quad \forall\, s \in [a,t] \qquad (3)$$

Since dF is a finite measure it follows that $g(\tau) \in L_1(dF;a,t)$; consequently $G(\tau)$, defined by $dG(\tau) = g(\tau)dF(\tau)$, is of bounded variation over $[a,t]$. It is possible to choose $G(a)=0$; then $G(s)$ is also bounded in $[a,t]$. Equation (3) becomes

$$\int_a^s h(s,\tau)dG(\tau) = 0 \quad , \quad \forall\, s \in [a,t] \qquad (4)$$

Equation (4) can be integrated by parts since both $G(\tau)$ and $h(s,\tau)$ are of bounded variation over $[a,t]$. The result is

$$h(s,s)G(s) - \int_a^s \frac{\partial h(s,\tau)}{\partial\tau} G(\tau)d\tau = 0 \quad , \quad \forall s\in[a,t]$$

Since $h(s,s)>\varepsilon>0$, it follows that

$$G(s) - \int_a^s \frac{G(\tau)}{h(s,s)} \cdot \frac{\partial h(s,\tau)}{\partial\tau}\, d\tau = 0 \quad , \quad \forall s\in[a,t] \qquad (5)$$

Equation (5) is a homogeneous Volterra equation of the second kind and it is known [12] that its unique, continuous L_2 solution is zero. Thus $G(s) \equiv 0$, which implies

$$g(\tau) = 0 \text{ a.e. } [dF]$$

and therefore $u=0$. ▌

A necessary condition is given by the following theorem.

Theorem 2: If $X(t) = \int_{-\infty}^{t} h(t,\tau)dz(\tau)$ and $H(X,t) = H(z,t)$, then $h(t,\tau)$ is not an analytic function at any neighborhood of a point on the diagonal $t=\tau$, unless at least one of the derivatives $\partial^{(i)}h(t,\tau)/\partial\tau^{(i)}|_{t=\tau}$, $i=0,1,2,\ldots$ is non-zero in that neighborhood.

Proof: If all the above derivatives are zero in a neighborhood of (t,t) then by the Taylor expansion $h(t,\tau)=0$ in that neighborhood; this implies that $h(t,\tau)$ is not proper [11] and therefore $H(X,t) \neq H(z,t)$. ▌

By applying Theorem 1 to the subclasses of time-invariant or separable systems, a well-known result can be obtained [13].

Part II: Conditions on the Process Itself

Let $X(t)\in C$. If $X(t)$ is wide sense stationary and pnd then it is known that $M=1$. Also if $X(t)$ is wide sense Markov, again $M=1$. On the other hand, there are well-known random processes that do not have unit multiplicity. For example the deterministic processes have $M=0$. The widely postulated band-limited processes belong to this category [14]. Finally Cramer [15] has given examples of pnd random processes with $M>1$. Since the autocovariance function $R(t,s)$ of the process determines its deterministic or pnd nature as well as its multiplicity, it is desirable to derive appropriate conditions on $R(t,s)$. However this seems to be rather difficult. Cramer* has given the following sufficient condition:

Proposition (Cramer): Let

$$X(t) = \sum_{i=1}^{M} \int_{-\infty}^{t} g_i(t,\tau)dz_i(\tau) \quad , \quad E[dz_i(\tau)]^2 = dF_i(\tau)$$

be the canonical decomposition of $X(t)$. If $g_i(t,t)>0$, $\forall t\in R$ and $\forall i$, and if both $g_i(t,\tau)$ and $\partial g_i(t,\tau)/\partial t$ are bounded and if the F_i's are absolutely continuous $(F_i'=f_i)$, then $M=1$.

This condition shows that, if a random process $X(t)\in C$ is to have multiplicity $M>1$, its components must exhibit some irregular behavior. From a similar line of reasoning a slightly different condition can be derived which shows the degree of irregularity required to be present in any finite interval and with respect to any two terms of the canonical decomposition. It is possible that the conditions given here may be relaxed.

Theorem 3: Let

$$X(t) = \sum_{i=1}^{M} \int_{-\infty}^{t} g_i(t,\tau)dz_i(\tau) \quad , \quad E[dz_i(\tau)]^2 = dF_i(\tau)$$

be the canonical decomposition of $X(t)$. There can be no two index values, $\varkappa$ and λ, such that $g_\lambda(t,\tau)$ has bounded first derivatives in $[a,t]$, $g_\varkappa(t,\tau)$ has bounded second derivatives in $[a,t]$, and $g_\varkappa(s,s)>\epsilon>0$, $\forall s\in[a,t]$, and such that $F_\varkappa(\tau)$ and $F_\lambda(\tau)$ are absolutely continuous in $[a,t]$, that is $F_\varkappa' = f_\varkappa$ and $F_\lambda' = f_\lambda$, with $f_\varkappa(\tau)$, $f_\lambda(\tau>\delta>0$.

Proof: If $M>1$, we shall show that there exists a $u\neq 0$ such that $u\in H(X,t)$ and $u\perp H(X,t)$ which is a contradiction. We have

---

*Lecture at Princeton University, March 1970.

$$u \in H(X,t) \Rightarrow u = \sum_{i=1}^{M} \int_{-\infty}^{t} h_i(\tau)\,dz_i(\tau) \qquad (6)$$

with $h_i(\tau) \in L_2(dF_i;-\infty,t)$, $i=1,\ldots,N$. On the other hand

$$u \perp H(X,t) \Rightarrow E[u \cdot X(s)] = 0 \quad , \quad \forall\ s \le t \qquad (7)$$

Equation (7) can be written

$$\sum_{i=1}^{M} \int_{-\infty}^{s} g_i(s,\tau)h_i(\tau)\,dF_i(\tau) = 0 \quad , \quad \forall\ s \le t \qquad (8)$$

Choose $h_i(\tau) \equiv 0$ for $i \neq \varkappa,\lambda$ and let

$$h_\lambda(\tau) = \begin{cases} \text{bounded in } [a,t] \\ 0 \quad , \quad \tau \notin [a,t] \end{cases}$$

Let $h_\varkappa(\tau)$ be undetermined in $[a,t]$ and be zero for $\tau \notin [a,t]$. Then Equation (8) becomes

$$\int_a^s g_\varkappa(s,\tau)h_\varkappa(\tau)\,dF_\varkappa(\tau) = -\int_a^s g_\lambda(s,\tau)h_\lambda(\tau)\,dF_\lambda(\tau), \forall s \in [a,t] \qquad (9)$$

The function $H_\varkappa(\tau)$ defined by $dH_\varkappa(\tau) = h_\varkappa(\tau)dF_\varkappa(\tau)$ is absolutely continuous in $[a,t]$ and therefore bounded and of bounded variation. Choose $H_\varkappa(a)=0$. Then, by integrating the left side of Equation (9) by parts, we obtain

$$g_\varkappa(s,s)H_\varkappa(s) - \int_a^s \frac{\partial g_\varkappa(s,\tau)}{\partial \tau} H_\varkappa(\tau)\,d\tau = -\int_a^s g_\lambda(s,\tau)h_\lambda(\tau)\,dF_\lambda(\tau)$$

Since $g_\varkappa(s,s) > \delta > 0$, we have

$$H_\varkappa(s) - \int_a^s \frac{\partial g_\varkappa(s,\tau)}{\partial \tau} \cdot \frac{H_\varkappa(\tau)}{g_\varkappa(s,s)}\,d\tau = \Omega(s) \qquad (10)$$

where

$$\Omega(s) = -\int_a^s \frac{g_\lambda(s,\tau)}{g_\varkappa(s,s)} h_\lambda(\tau)\,dF_\lambda(\tau)$$

Equation (10) is a Volterra equation of the second kind; $\Omega(s)$ is bounded and non-zero in $[a,t]$; therefore there exists a unique non-zero bounded solution $H_\varkappa(s)$. It remains to be shown that $H_\varkappa(s)$ is absolutely continuous and that $h_\varkappa(s) = dH_\varkappa(s)/dF_\varkappa(s)$ is in $L_2(dF_\varkappa;a,t)$. It can be proved that a function of the form $\int_a^s G(s,\tau)h(\tau)d\tau$ is absolutely continuous if G has bounded derivatives and $\int_a^t\int [\partial G(s,\tau)/\partial s]\ h(\tau)d\tau ds < \infty$. These conditions are satisfied for both $\Omega(s)$ and the second term of the left side of Equation (10). Therefore $H_\varkappa(s)$ is indeed absolutely continuous. On differentiating Equation (10) it can be seen that $H'_\varkappa(s)$ is bounded from the properties of $g_\varkappa(s,\tau)$ and $g_\lambda(s,\tau)$. But $H'_\varkappa(s) = f_\varkappa(s)h_\varkappa(s)$; therefore

$$h_\varkappa(s) < \frac{H'_\varkappa(s)}{\delta}$$

Since $h_\varkappa(s)$ is bounded and $dF_\varkappa$ is a finite measure $h_\varkappa(s) \in L_2(dF_\varkappa; a,t)$. Therefore $u \neq 0$. ∎

IV. PREWHITENING IN LINEAR LEAST-MEAN-SQUARE ESTIMATION

As discussed earlier, in order for the prewhitening method to be valid, it is necessary and sufficient that the observation process $Y(t)$ be pnd with multiplicity $M=1$. In the following, some statements will be proved relating the multiplicity of $Y(t)$ to the properties of the signal and noise processes $Z(t)$ and $N(t)$. Throughout we shall assume that $H(Y,t)=H(Z,t)\oplus H(N,t)$, $\forall t\in R$. This assumption is rather severe for the estimation problem. It implies that zero-error estimation of the signal is theoretically possible. However there are cases for which it holds and furthermore the results are valid in problems of representation of multivariate random processes. In general we have $H(Y,t)\subset H(Z,t)\oplus H(N,t)$.

Preliminary Remarks

1. From $Y(t)=Z(t)+N(t)$, $\forall t\in R$, and $H(Z)\perp H(N)$ it follows that the projection of $Y(t)$ on $H(Z)$, denoted by $P[Y(t)]$, is equal to $Z(t)$. Therefore $P[H(Y,t)] = H(Z,t)$. Furthermore, if $\Delta=[s,t)$ and if $H(Y,\Delta)$ denotes the orthogonal complement of $H(Y,s)$ with respect to $H(Y,t)$, then $P[H(Y,\Delta)] = H(Z,\Delta)$.

2. Given that $y(t)$ is a left continuous process in C with orthogonal increments, a finite orthogonal random measure ψ can be defined on the semi-algebra of right-open, left-closed intervals $\Delta=[s,t)$ by $\psi(\Delta) = y(t)-y(s)$. This measure can be extended to the σ-algebra of Borel sets [16]. The same holds for the associated real measure defined by $F(\Delta)=E|\psi(\Delta)|^2$. Furthermore the space $H(y,\Delta)$ which is spanned by $\psi(\delta)$ for all $\delta\in\Delta$ can also be extended to the Borel field as an abstract measure. If D is a Borel set, then $H(Y,D)$ is the space spanned by $\psi(d)$ for all Borel subsets $d\in D$.

3. The family of all finite real measures defined on the Borel field forms a lattice ordered by absolute continuity. Given two measures F_1, F_2 we can decompose F_2 as $F_2'+F_2''$ with $F_2'\perp F_2''$, $F_2'<<F_1$, and $F_2''\perp F_1$ by the Lebesgue decomposition theorem [17]. The least upper bound (lub) of F_1 and F_2 is $F=F_1+F_2''$ and their greatest lower bound (glb) is $f=F_2'$. It is interesting to notice that if $y_1(t)$ and $y_2(t)$ are two mutually orthogonal processes in C with orthogonal increments, with corresponding real measures F_1 and F_2, then the space $H(y_1,t)\oplus H(y_2,t)$ is equal to $H(y_1',t)\oplus H(y_2',t)$, $\forall t\in R$, where $y_1'(t)$ and $y_2'(t)$ are again two mutually orthogonal processes in C with orthogonal increments with respective measures $F=F_1\vee F_2$ and $f=F_1\wedge F_2$.

Principal Results and Proofs

Lemma 1: If $Y(t)=Z(t)+N(t)$ and $H(Z)\perp H(N)$, then $Y(t)$ is purely non-deterministic (pnd) iff both $Z(t)$ and $N(t)$ are pnd.
Proof: From $H(Y,t) = H(Z,t)\oplus H(N,t)$, $\forall t\in R$ it follows that $H(Y,-\infty) = H(Z,-\infty)\oplus H(N,-\infty)$ and therefore $H(Y,-\infty) = 0$ iff $H(Z,-\infty)=H(N,-\infty) = 0$. ∎

Theorem 1: If, in the situation of Lemma 1, Z(t) and N(t) are pnd with multiplicities M_1 and M_2 respectively, then the multiplicity M of $Y(t) = Z(t)+N(t)$ satisfies

$$\max[M_1,M_2] \le M \le M_1+M_2 \qquad (11)$$

Proof: Let B be a minimal generating subspace for Y(t). Then $H(Y,\Delta) = \underset{\delta\in\Delta}{G}\{P_Y(\delta)[B]\}$, $\forall\Delta=[s,t)$. By Remark 1 we have

$$H(Z,\Delta) = P[H(Y,\Delta)] = \underset{\delta\in\Delta}{G}\{P\cdot P_Y(\delta)[B]\}$$

But $P\cdot P_Y(\delta) = P_Z(\delta)$ and therefore

$$H(Z,\Delta) = \underset{\delta\in\Delta}{G}\{P_Z(\delta)[B]\} = \underset{\delta\in\Delta}{G}\{P_Z(\delta)\cdot P[B]\}$$

This means that $B'=P[B]$ is a generating subspace for Z(t). The dimension of B' is $M'\le M$; consequently $M_1\le M$. Similarly we can prove that $M_2\le M$ and therefore that the left side of the inequality of Eq. (11) is true.

If B_1 and B_2 are minimal generating subspaces for Z(t) and N(t) respectively, then $B_1\oplus B_2$ is a generating subspace for H(Y,t), since $P_Y(\delta) = P_Z(\delta)+P_N(\delta)$. Therefore $M\le M_1+M_2$. ▌

Corollary 1: In order that Y(t) be a pnd process with multiplicity one, it is necessary that Z(t) and N(t) each have multiplicity one.

Assuming now that $M_1=M_2=1$ it is of interest to specify when M=1. The following theorem provides a necessary and sufficient condition.

Theorem 2: Let $y_1(t)$ and $y_2(t)$ be mutually orthogonal, pnd random processes in C with orthogonal increments. Let F_1 and F_2 be the corresponding real measures. Then the multiplicity M of $H(y_1,t)\oplus H(y_2,t) = H(y_0,t)$ is equal to unity iff F_1 and F_2 are mutually singular.

Proof: First we note that $M\le 2$ as was proved in the preceding theorem and in [5].

(i) Sufficiency: If $F_1\perp F_2$, then by the construction explained in Remark 3 it is obvious that M=1, since f=0.

(ii) Necessity: If M=1, let F_0 be the real measure associated with $H(y_0,t)$. If $F_0(D)=0$ for some Borel set D it follows by Remark 2 that $H(y_0,D)=0$. Therefore $H(y_1,D)=H(y_2,D)=0$ and $F_1(D)=F_2(D)=0$. Consequently $F_1<<F_0$ and $F_2<<F_0$. This implies [16] that ψ_1 and ψ_2 have Radon-Nikodym derivatives with respect to ψ_0; that is

$$\psi_i(D) = \int_D g_i(\tau)\,\psi_0(d\tau) \quad , \quad i=1,2$$

But ψ_1 and ψ_2 are mutually orthogonal; therefore

$$E[\psi_1(D)\psi_2(D)] = \int_D g_1(\tau)g_2(\tau)F_0(d\tau) = 0$$

for every Borel set D; therefore

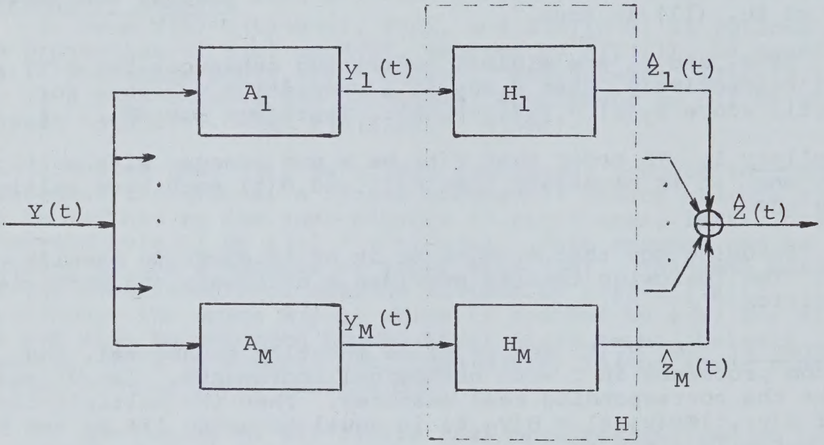
$$g_1(\tau)g_2(\tau) = 0 \text{ a.e.} [F_0]$$

Then F_1 and F_2 have disjoint supports, which means that $F_1 \perp F_2$. ∎

A direct extension of this theorem is the following:

Theorem 3: Let $y_1(t),\ldots,y_M(t)$ be mutually orthogonal, pnd random processes in C with orthogonal increments. The multiplicity of $H(Y,t) = H(y_1,t)\oplus\ldots\oplus H(y_M,t)$ is less than M iff the glb of the measures $F_1,\ldots,F_M$ is the null measure.

In case $M>1$, multiple or parallel prewhitening can be applied as shown in the following diagram.



If the initial assumption that $H(Y,t) = H(Z,t)\oplus H(N,t)$, $\forall t \in R$, is omitted, it is not clear that all the previous results hold. If N(t) is assumed to be a process with orthogonal increments, then it can be shown that the process $y(t) = Y(t)-P_Y^t[Z(t)]$ has orthogonal increments. The problem is then to determine under what conditions $H(Y,t) \subset H(y,t)$. A complete answer to this question as well as to the problem of specifying when a given random process has unit multiplicity have not yet been obtained.

ACKNOWLEDGMENT

The authors wish to acknowledge the support of the Air Force Office of Scientific Research under AFOSR Grant 1333-67, of the National Science Foundation under Grant GK-1439, and of the U.S. Army Research Office under Contract DAHCO4-69-C-0012.

REFERENCES

1. H. Cramer, "On the Structure of Purely Non-deterministic Stochastic Processes", Arkiv för Matematik, Vol. 4, pp. 249-266, 1961.

2. H. Cramer, "On Some Classes of Non-stationary Stochastic Processes", Proc. of the Fourth Berkeley Symposium on Statistics and Applied Probability, Berkeley and Los Angeles, University of California Press, Vol. II, pp. 57-78, 1960.

3. T. Hida, "Canonical Representations of Gaussian Processes and Their Applications", Memoirs of the College of Science, University of Kyoto, Ser. A, Vol. XXXIII, Math. No. 1, 1960.

4. G. Kallianpur, V. Mandrekar, "On the Connection between Multiplicity Theory and O. Hanner's Time Domain Analysis of Weakly Stationary Stochastic Processes", Technical Report No. 49, University of Minnesota, Dept. of Statistics, 1964.

5. V. Mandrekar, "On Multivariate Wide Sense Markov Processes", Nagoya Mathematical Journal, Vol. 33, pp. 7-19, 1968.

6. N.I. Akhiezer, I.M. Glazman, Theory of Linear Operators in Hilbert Space, Frederick Ungar Publ. Co., 1963.

7. J.L. Doob, Stochastic Processes, J. Wiley, 1953.

8. H. Bode, C. Shannon, "A Simplified Derivation of Linear Least Square Smoothing and Prediction Theory", Proc. of the I.R.E., Vol. 38, pp. 417-425, 1950.

9. T. Kailath, "An Innovations Approach to Least-Squares Estimation Part I: Linear Filtering in Additive White Noise", IEEE Transactions Automatic Control, Vol. AC-13, No. 6, December 1968.

10. T. Kailath, "Likelihood Ratios for Gaussain Processes", IEEE Trans. Information Theory, Vol. IT-16, No. 3, May 1970.

11. P. Levy, "Sur une Classe de Courbes de l'Espace de Hilbert et sur une Equation Integrale Non-lineaire", Ann. Sci. Ecole Normale Superieure, t. 73, pp. 121-156, 1956.

12. F. Smithies, Integral Equations, Cambridge University Press, 1958.

13. E.B. Stear, "Shaping Filters for Stochastic Processes", Modern Control Systems Theory, Edited by C.T. Leondes, McGraw-Hill, 1965.

14. Y.K. Belyaev, "Analytic Random Processes", Theory of Probability and its Applications, p. 402, 1959.

15. H. Cramer, "A Contribution to the Multiplicity Theory of Stochastic Processes", Proc. of the Fifth Berkeley Symposium on Statistics and Applied Probability, Berkeley and Los Angeles, University of California Press, Vol. 2, pp. 215-221, 1965.

16. S. Cambanis, "On Hilbert Space Valued Measures", Technical Report No. 29, Princeton University, Dept. of Electrical Engineering, July 1969.

17. H.L. Royden, Real Analysis, McMillan, New York, 1963.

SOME CLASSES OF ANALYTIC RANDOM PROCESSES

KEVIN C. DALY
Department of the Air Force
Air Force Unit Post Office
Los Angeles, Cal. 90045

JOHN B. THOMAS
Department of Elec. Eng.
Princeton University
Princeton, N.J. 08540

I. INTRODUCTION

The functional properties of realizations of random processes are of considerable theoretical and practical interest. These properties are particularly important when processes which are well-defined mathematically are considered as models for natural random phenomena such as interfering noise sources or variations in system parameters. In such cases the properties of the individual realizations or sample functions of the processes must be compatible with their physical origin. Investigation of the sample function behavior of the modeling processes can often be valuable not only as an aid in choosing the appropriate mathematical model but also for indicating those areas in which the model may be inappropriate.

The concern of this paper will not be with physical processes per se but with the functional properties of many common mathematical approximations to such processes.

II. BACKGROUND

We will in general consider complex processes $x(t,\omega)$. When there is no possibility for confusion the ω variable will be omitted and the processes written $x(t)$. In order to discuss sample function properties in a meaningful way and to insure that the processes are uniquely defined by the class of finite order distributions $P_{x_1,x_2,\ldots,x_j}(A_1,A_2,\ldots,A_j) = \text{Prob}[x(t_1)\in A_1, x(t_2)\in A_2,\ldots,x(t_j)\in A_j]$, we will require that the processes under consideration be stochastically continuous and separable (in this case, with respect to two-dimensional Borel sets). The full implications of the separability restriction are discussed in Doob [1].

Unless otherwise noted all processes will be assumed to be zero mean; that is, $E\{x(t)\}\equiv 0$. This does not restrict the class of processes to be considered since any process $y(t)$ may be expressed as $y(t)=x(t)+m(t)$ where $m(t)=E\{y(t)\}$. We will often consider the correlation function of the process which will be written

$$R_x(t,s) = E\{x(t)\overline{x}(s)\}$$

For processes which are at least wide sense stationary, this last expression becomes

$$R_x(\tau) = R_x(t,t+\tau)$$

Among the sample function properties to be considered are continuity, boundedness, and differentiability. We will say that a random process $x(t,\omega)$ possesses an analytic property P in an interval T if the sample functions of this process possess the property P in T with probability one (w.p.1). Possession of the property P in mean square (m.s.) or in probability (i.p.) are not of direct interest here since these do not necessarily imply anything about the sample function behavior of the process.

A sufficient condition for continuity of the sample functions of a random process (w.p.1) at the point t' is

$$R_x(t'+\tau,t'+\tau)-R_x(t'+\tau,t')-R_x(t',t'+\tau)+R_x(t',t') = 0\left(\frac{|\tau|}{|\log|\tau||^q}\right), \tau\to 0 \qquad (1)$$

for some $q>3$. This condition, originally due to Kolmogorov, is proved in Loeve [2]. Eq. (1) will be satisfied if, for example

$$\left| \left. \frac{\partial^2 R_x(t,s)}{\partial t \partial s} \right|_{t=s=t'} \right| < \infty \qquad (2)$$

For stationary processes, Eq. (2) reduces to

$$\left| \left. \frac{d^2 R_x(\tau)}{d^2 \tau} \right|_{\tau=0} \right| < \infty \qquad (2a)$$

Equivalent sufficient conditions can be deduced for the existence of continuous derivatives of the sample functions of x(t). In particular, if

$$R_x(t'+2\tau,t'+2\tau)-2R_x(t'+2\tau,t'+\tau)-2R_x(t'+\tau,t'+2\tau)+R_x(t'+2\tau,t')$$

$$+4R_x(t'+\tau,t'+\tau)+R_x(t',t'+2\tau)-2R_x(t'+\tau,t')-2R_x(t',t'+\tau)+R_x(t',t')$$

$$= 0\left(\frac{|\tau|^3}{|\log|\tau||^q}\right), \quad \tau\to 0 \qquad (3)$$

for some $q>3$, the sample functions of the process x(t) will have continuous derivatives (w.p.1). Eq. (3) will be satisfied if

$$\left| \left. \frac{\partial^4 R_x(t,s)}{\partial^2 t \, \partial^2 s} \right|_{t=s=t'} \right| < \infty \qquad (4)$$

The obvious equivalent condition for stationary processes is

$$\left| \left. \frac{d^4 R_x(\tau)}{d\tau^4} \right|_{\tau=0} \right| < \infty \qquad (4a)$$

The same method of proof may be used to demonstrate the following theorem.

Theorem: A sufficient condition for a random process $x(t)$ to possess n derivatives in some neighborhood of the point t' (that is, for $|t-t'|<r$) is for the correlation function $R_x(t,s)$ of the process to be (n+1) times differentiable for $|t-t'|<r$, $|s-t'|<r$; that is, for

$$\left| \left. \frac{\partial^{2n+2} R_x(t,s)}{\partial^{n+1} t\, \partial^{n+1} s} \right|_{t=s=t'} \right| < \infty \quad , \quad \begin{matrix} |t-t'|<r \\ |s-t'|<r \end{matrix} \tag{5}$$

The above sufficient conditions are true for any random process. In the special case of a wide sense stationary Gaussian process, much more information may be obtained on the sample function behavior of the process. Since, under the assumption of zero mean, a Gaussian random process is completely determined by its autocorrelation function, it should not be surprising that many sample function restrictions may be translated into conditions on this autocorrelation function or, equivalently, on its Fourier transform, the power spectral density. The weakest sufficient condition for sample function continuity of Gaussian random processes is that of Hunt [3].

a) If $x(t)$ is a stationary Gaussian random process with power spectral density $\phi(\lambda)$ and if

$$\int_0^\infty [\log(1+\lambda)]^a \, \phi(\lambda) d\lambda < \infty \tag{6}$$

for some $a>1$, then the sample functions of the process $x(t)$ are continuous (w.p.1).

b) If instead of Eq. (6) we have

$$\int_0^\infty \lambda^2 [\log(1+\lambda)]^a \, \phi(\lambda) d\lambda < \infty \tag{7}$$

for some $a>1$, then the sample functions of $x(t)$ have continuous derivatives (w.p.1).

From a result of Belayev [4], Eq. (6) is equivalent to the condition that

$$R_x(0) - R_x(\tau) \le \frac{C}{|\log|\tau||^b} \tag{8}$$

for some $b>1$, $C>0$ and for all sufficiently small τ. That Eq. (6) is a weaker condition than Eq. (2a) can be seen from the following example.

Consider the Gauss-Markov process; for this Gaussian process $\phi(\lambda) = 1/(1+\lambda^2)$. Let a=2 and consider

$$I = \int_0^\infty [\log(1+\lambda)]^2 \frac{1}{1+\lambda^2} d\lambda$$

$$= \int_0^1 [\log(1+\lambda)]^2 \frac{1}{1+\lambda^2}\, d\lambda + \int_1^\infty [\log(1+\lambda)]^2 \frac{1}{1+\lambda^2}\, d\lambda$$

$$\le \int_0^1 [\log(2)]^2 \frac{1}{1+\lambda^2}\, d\lambda + \int_1^\infty [\log(2\lambda)]^2 \frac{1}{1+\lambda^2}\, d\lambda$$

$$= \int_0^\infty [\log(2)]^2 \frac{1}{1+\lambda^2}\, d\lambda + \int_1^\infty \frac{\log^2(\lambda)}{1+\lambda^2}\, d\lambda + 2\log 2 \int_1^\infty \frac{\log\lambda}{1+\lambda^2}\, d\lambda$$

$$= c^2 \int_0^\infty \frac{1}{1+\lambda^2}\, d\lambda + \int_0^1 \frac{\log^2(\omega)}{1+\omega^2}\, d\omega - 2c \int_0^1 \frac{\log(\omega)}{1+\omega^2}\, d\omega$$

$$= c^2 \frac{\pi}{2} + \frac{\pi^3}{16} + 2cG < 4$$

where $c=\log 2=0.6931\ldots$ and G=Catalan's Constant=0.9159... . Since the integrand is always positive

$$0 < I < 4$$

and by Eq. (6) the Gauss-Markov process has continuous sample functions (w.p.1).

The sufficient conditions for sample function continuity of Gaussian processes are much weaker than those for random processes in general. Although necessary conditions for sample function continuity have not been found, stationary Gaussian processes have been demonstrated with correlation functions satisfying

$$R_x(0) - R_x(\tau) \le \frac{C}{\log|\tau| \prod_{j=2}^{n} \log_j^2 |\tau|} \tag{9}$$

with $C>0$ for any n and for all τ sufficiently small. In this expression $\log_j(\tau)$ is the iterated logarithm

$$\log_j(\tau) = \underbrace{\log\log\log\ldots\log}_{j\text{ times}}(\tau) \tag{10}$$

At the same time, the process has discontinuous sample functions (w.p.1). [See Fernique (5)]. Thus, for Gaussian processes Eq. (6) (or equivalently Eq. (8)) may be very close to a necessary condition.

Although continuity would appear to be a reasonable requirement to place on random processes, there are valid examples of random process models for which continuity is not required. The Poisson process, for example, which is often used as a model of queueing processes has sample functions which are discontinuous (w.p.1). The sample functions of the Poisson process have a countable number of isolated discontinuities of the first kind (i.e., right and left hand limits exist at the point of discontinuity but are not equal). It follows from the work of Dobrushin and Belayev that this type of behavior is not possible for Gaussian processes. If a Gaussian random process does not have continuous sample functions (w.p.1), then with probability one the process has a discontinuity of the second kind at every point t′ [Dobrushin [6]] and with probability one the process is unbounded in all finite intervals [Belayev [4]]. Gaussian processes then are either very well behaved or extremely poorly behaved from a physical point of view.

III. ANALYTIC RANDOM PROCESSES

By utilizing conditions such as Eq. (6), classes of random processes may be defined with successively more regular sample function properties. For example, random processes whose power spectral density is a rational function of the form

$$\varphi(\omega) = \frac{\sum_{j=1}^{n} \alpha_j \omega^{2j}}{\sum_{k=1}^{m} \beta_k \omega^{2k}} , \quad m>n \tag{11}$$

are guaranteed (w.p.1) to have derivatives of at least order m-n-1. A case of particular interest is when $m-n=\infty$, that is when it would appear logical to conclude that the sample functions of the process possess an infinite number of derivatives or that they are analytic functions (of t) with probability one. This type of process, for example, might correspond to white noise passed through an infinite number of isolated low pass filters. The class of analytic random processes [processes whose sample functions are continuous (w.p.1)] is, in fact, much wider than the class discussed above. These processes are of interest both because of their sample function behavior and because they include many processes commonly used to approximate large classes of random phenomena.

We will first prove a sufficient condition for a random process x(t) to be an analytic random process:

Theorem: If the correlation function $R_x(t,s)$ of the process x(t) is an analytic function of the two variables t and s in a neighborhood of the point (t_0,t_0), then the sample functions of the process are analytic functions of t (w.p.1) in a neighborhood of the point t_0.

Proof: Since $R_x(t,s)$ is an analytic function of t and s in a neighborhood of the point (t_0,t_0) there is some $r>0$ such that

$$R_x(t,s) = \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \left. \frac{\partial^{k+m} R_x(t,s)}{\partial t^k \, \partial s^m} \right|_{t=s=t_0} \frac{(t-t_0)^k (s-t_0)^m}{k! \; m!} \tag{12}$$

for $|t-t_0|<r$ and $|s-t_0|<r$. From the condition of Eq. (6), the quantity

$$\frac{d^n}{dt^n}\, x(t) = x^{(n)}(t)$$

exists for all n for $|t-t_0|<r$. Consider the processes

$$x_n(t) = \sum_{k=0}^{n} x^{(k)}(t_0)\, \frac{(t-t_0)^k}{k!} \tag{13}$$

These $x_n(t)$ processes are analytic functions (polynomials, in fact) of t. Let

$$\begin{aligned} D_n &= E\{|x(t)-x_n(t)|^2\} \\ &= R_x(t,t) - 2\sum_{k=0}^{n} \frac{\partial^k}{\partial s^k} R_x(t,s)\Big|_{s=t_0} \frac{(t-t_0)^k}{k!} \\ &\quad + \sum_{k=0}^{n}\sum_{m=0}^{n} \frac{\partial^{k+m}}{\partial t^k\, \partial s^m} R_x(t,s)\Big|_{t=s=t_0} \frac{(t-t_0)^{k+m}}{k!\ m!} \\ &= \sum_{k=0}^{\infty}\sum_{m=0}^{\infty} \frac{\partial^{k+m}}{\partial t^k\, \partial s^m} R_x(t,s)\Big|_{t=s=t_0} \frac{(t-t_0)^{k+m}}{k!\ m!} \\ &\quad + \sum_{k=0}^{n}\sum_{m=0}^{n} \frac{\partial^{k+m}}{\partial t^k\, \partial s^m} R_x(t,s)\Big|_{t=s=t_0} \frac{(t-t_0)^{k+m}}{k!\ m!} \\ &\quad -2\sum_{k=0}^{\infty}\sum_{m=0}^{n} \frac{\partial^{k+m}}{\partial t^k\, \partial s^m} R_x(t,s)\Big|_{t=s=t_0} \frac{(t-t_0)^{k+m}}{k!\ m!} \end{aligned} \tag{14}$$

The convergence of the three summations as $n\to\infty$ for $|t-t_0|<r$ is guaranteed by the analyticity of $R_x(t,s)$, and therefore

$$\lim_{n\to\infty} D_n = R_x(t,t)+R_x(t,t)-2R_x(t,t) = 0$$

Thus

$$\underset{n\to\infty}{\text{l.i.m.}}\; x_n(t) = x(t) \tag{15}$$

The second part of the proof consists of showing that $\lim_{n\to\infty} x_n(t)$ exists (w.p.1), or equivalently, showing that with

$$d_n(t) = \sum_{k=n}^{\infty} x^{(n)}(t_0)\, \frac{(t-t_0)^k}{k!} \quad , \quad |t-t_0|<r$$

we have

$$\lim_{n\to\infty} d_n(t) = 0 \quad \text{(w.p.1)} \qquad (16)$$

Eq. (16) will be satisfied if

$$S = \sum_{k=0}^{\infty} E\{d_n^2(t)\} < \infty \quad , \quad |t-t_0|<r \qquad (17)$$

But

$$E\{d_n^2(t)\} = \sum_{k=n}^{\infty}\sum_{m=n}^{\infty} \frac{\partial^{k+m}R_x(t,s)}{\partial t^k\ \partial s^m}\bigg|_{t=s=t_0} \frac{(t-t_0)^{k+m}}{k!\ m!}$$

Thus

$$S = \sum_{k=0}^{\infty} E\{d_n^2(t)\} = \sum_{j=0}^{\infty}\left[\sum_{k=j}^{\infty}\sum_{m=j}^{\infty} \frac{\partial^{k+m}R_x(t,s)}{\partial t^k\ \partial s^m}\bigg|_{t=s=t_0} \frac{(t-t_0)^{k+m}}{k!\ m!}\right]$$

$$= \sum_{k=0}^{\infty}\sum_{m=0}^{\infty} \frac{b(k,m)\,\partial^{k+m}R_x(t,s)}{\partial t^k\ \partial s^m}\bigg|_{t=s=t_0} \frac{(t-t_0)^{k+m}}{k!\ m!}$$

where $b(k,m)=1+\min(k,m)$. Thus

$$S \le \sum_{k=0}^{\infty}\sum_{m=0}^{\infty} \frac{\partial^{k+m}R_x(t,s)}{\partial t^k\ \partial s^m}\bigg|_{t=s=t_0} \frac{(t-t_0)^{k+m}}{k!\ m!}$$

$$+ \sum_{k=0}^{\infty}\sum_{m=0}^{\infty} \frac{\partial^{k+m}R_x(t,s)}{\partial t^k\ \partial s^m} \frac{k\cdot m(t-t_0)^{k+m}}{k!\ m!} < \infty$$

The convergence of the first sum follows from the analyticity of $R_x(t,s)$; the convergence of the second sum follows from the convergence of the power series for $\partial^2 R_x(t,s)/\partial t\partial s$ where $|t-t_0|<r$ and $|s-t_0|<r$. Thus $\hat{x}(t)=\lim x_n(t)$ exists (w.p.1) and, therefore, is an analytic random process so that

$$x(t) = \hat{x}(t) \quad \text{(w.p.1)} \quad \text{for } |t-t_0|<r \qquad \blacksquare$$

It has been shown by Belayev [7] that, for Gaussian processes, the conditions of the theorem are also necessary.

For stationary processes the conditions of the theorem reduce to the condition that $R_x(\tau)$ be an analytic function of τ; evidently $R_x(\tau)$ is an analytic function of τ for $|\tau|<r$ if

$$\int_{-\infty}^{\infty} e^{\sigma\omega} \phi(\omega)d\omega < \infty \qquad |\sigma|\leq r \tag{18}$$

It should be noted that the convergence of the integral in Eq. (18) implies the divergence of the integral

$$I = \int_{-\infty}^{\infty} \frac{\log[\phi(\omega)]}{1+\omega^2} d\omega \tag{19}$$

and thus all processes satisfying Eq. (18) are completely deterministic. In particular, stationary analytic Gaussian processes are completely deterministic. On the other hand Gaussian processes with rational spectra are completely non-deterministic.

IV. EXAMPLES

There are several classes of analytic random processes which are of particular interest; among these are random almost periodic functions, random processes with rational correlation, and bandlimited random processes.

Random Almost Periodic Functions: Random processes of the form

$$x(t) = \sum_{k=0}^{\infty} \alpha_k \cos(\omega_k t+\beta_k)$$

where

$$\sum_{k=0}^{\infty} E\{\alpha_k^2\} < \infty$$

and the β_k are independent identically distributed random variables uniformly distributed on $[0,2\pi]$ will be called random almost periodic functions. In the case where $\omega_k=k\omega_0$ the processes will be called random Fourier series. These processes are often used to approximate more general random processes "in a mean square sense". However, they have some unique sample function properties. As defined, they are stationary but not ergodic. In the case of Gaussian processes, the Gaussian random almost periodic functions are the only Gaussian random processes for which the sample functions are bounded on the whole real line.

Random Processes with Rational Correlation: These processes are, in a sense, analogous to random processes with rational spectra. However the sample function properties of the two classes of processes are radically different. Although time does not permit a full development of the properties of these processes, investigation indicates that large classes of random processes x(t) may be approximated by

$$x(t) = x_c(t) + x_s(t)$$

where $x_c(t)$ is a process with a rational correlation function and $x_s(t)$ is a process with a rational spectrum. In the limit (as the order of the approximation increases without bound) $x_c(t)$ and $x_s(t)$ correspond to the deterministic and purely nondeterministic components of the processes, respectively, in the Cramer decomposition [8].

Bandlimited Random Processes: Convergence (w.p.1) of the sampling representation to the sample functions of bandlimited random processes has been shown by several authors [7,9]; that is, if

$$R_x(\tau) = \frac{1}{2\pi} \int_{-W}^{W} e^{j\tau\omega} \phi(\omega) d\omega$$

then

$$x(t) = \sum_{k=-\infty}^{\infty} x\left(\frac{k\pi}{W}\right) \frac{\sin[W(t-k\pi/W)]}{W(t-k\pi/W)} \quad (\text{w.p.1}) \tag{20}$$

The sum on the r.h.s. of Eq. (20) is an analytic function of t. Many properties of the sample functions of bandlimited random processes may be derived from Eq. (20). Several of these will be discussed in the final section.

V. RANDOM FUNCTIONS OF A COMPLEX VARIABLE

It can be shown easily from the theorem of Section III that a sufficient condition for the sample functions of a stationary process to be analytic in a strip $|I_m z|<r$, $z=t+j\sigma$, is for the correlation of the process to be analytic for $|\tau|<r$; that is, for

$$\int_{-\infty}^{\infty} e^{r\omega} \phi(\omega) d\omega < \infty \tag{21}$$

Thus, for example, processes with spectrum $\phi(\omega)=e^{-\alpha|\omega|}$ have sample functions (w.p.1) which may be extended to analytic processes in the strip $|I_m z|<\alpha$. Bandlimited random processes have sample functions which may be extended to analytic functions in the entire finite complex plane; that is, sample functions of bandlimited random processes may be considered (w.p.1) as restrictions of entire functions to the real line. It is well known that such functions are completely determined by their zeros.

From the representation of Eq. (20), sample functions of bandlimited random processes are exponential entire functions of type W. For functions of this type, the following has been shown to be true [10]:

Let n(r) be the number of zeros of the function x(t) occurring in a square of side r in the complex plane; then

$$\lim_{r\to\infty} \frac{n(r)}{r} = \frac{W}{\pi} \tag{22}$$

In other words, the zeros of the sample functions of bandlimited random processes (considered as functions of a complex variable $z=t+js$) occur at the Nyquist rate. As a degenerate case, consider the example of the random process $x(t)=\cos(Wt+\alpha)$, α uniformly distributed $[0,2\pi]$. The complex process $x(z)=x(t+js)$ has all its zeros on the real line and they obviously occur at a rate W/π. As a second example, we might consider bandlimited Gaussian random processes. Rice [11] has shown that for these processes, real zeros occur at a rate

$$r_r = \left[\frac{\int_{-W}^{W} \omega^2 \; \phi(\omega)d\omega}{\int_{-W}^{W} \phi(\omega)d\omega}\right]^{1/2} \frac{1}{\pi} \tag{23}$$

where $r_r \leq W/\pi$ and $r_r = W/\pi$ iff $\phi(\omega) = \frac{1}{2}[\delta(\omega+W)+\delta(\omega-W)]$; that is, if $x(t)=a\cos(Wt+\alpha)$. The remaining zeros of these processes occur for complex values of $z=t+js$ and are not available in the "real world". The question remains of the location of these zeros in the complex plane and whether there is any method for bringing them to the real line so that they will be observable in real time.

The zeros of $x(z)$ are not uniformly distributed in the complex plane; in fact, if the zeros of the process are denoted by

$$z_n = r_n e^{j\theta_n} \tag{24}$$

then

$$\sum_{n=1}^{\infty} \frac{\sin\theta_n}{r_n}$$

is absolutely convergent; that is, most of the zeros of the process occur within an angular region of the real axis. In the case of real random processes [processes for which $x(z)$ is real for real z] the zeros of the derivative (with respect to t) of the process must occur at a rate which is at least equal to the rate of the zeros of $x(t)$ since $x(t)$ must have at least one inflection point between any two zeros. The real zeros of the process $x'(z)=\partial x(z)/\partial t$ thus occur at a rate which is at least as great as that of the real zeros of $x(z)$. Thus differentiation brings complex zeros down to the real line. It has been found, for example, that, when vocal signals are passed through a nonlinear zero-memory filter for which $q(0)=0$, the signals are more intelligible if they are "differentiated" before and integrated after passing through the distorting medium.

The preceding examples have illustrated a few of the results of the theory of complex variables which may be applied to analytic random processes.

ACKNOWLEDGMENT

The authors wish to acknowledge the support of the Air Force Office of Scientific Research under AFOSR Grant 1333-67, of the National Science Foundation under Grant GK-1439, and of the U.S. Army Research Office under Contract DAHCO4-69-C-0012.

REFERENCES

1. J.L. Doob, Stochastic Processes, Wiley, New York, 1953.

2. M. Loeve, Probability Theory, Van Nostrand, Princeton, 1960.

3. G.A. Hunt, "Random Fourier Transforms", Transactions of the American Mathematical Society, Vol. 71 (1951), pp. 38-69.

4. Y.K. Belayev, "Continuity and Hölder's Conditions for Sample Functions of Stationary Gaussian Processes", Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, pp. 23-33.

5. X. Fernique, "Continuitė des Processus Gaussiens", Compes Rendus Academy Sci., Paris, Vol. 258, pp. 6058-6060.

6. R.I. Dobrushin, "Properties of Sample Functions of a Stationary Gaussian Process", Theory of Probability and its Applications, Vol. 5, pp. 120-122.

7. Y.K. Belayev, "Analytic Random Processes", Theory of Probability and its Applications, Vol. 4, pp. 402-409.

8. H. Cramer, "Dėcompositions Orthogonales du Certains Processus Stochastiques", Ann. Fac. Sci. Clermont, Vol. 11, pp. 15-21.

9. M. Zakai, "Bandlimited Functions and the Sampling Theorem", Information and Control, Vol. 8, pp. 143-158.

10. E.C. Titchmarsh, "The Zeros of Certain Integral Functions", Proceedings of the London Math. Society, Vol. 25, pp.283-302.

11. S.O. Rice, "Mathematical Analysis of Random Noise", Bell System Technical Journal, Vol. 23, pp. 282-332; Vol. 24, pp. 46-156.

ON THE REPRESENTATION OF WEAKLY CONTINUOUS STOCHASTIC PROCESSES

STAMATIS CAMBANIS[*]
University of North Carolina
Chapel Hill, North Carolina

ELIAS MASRY[**]
University of California
La Jolla, California

ABSTRACT

A novel approach to obtaining series representations in the stochastic mean for weakly, and therefore mean square, continuous stochastic processes is presented. Two distinct orthogonal series representations are derived for the entire class of weakly continuous stochastic processes over a broad class of Lebesgue sets of the real line, and a constructive procedure to obtain them explicitly is given. They include as particular cases all earlier representations. Also they are shown to converge almost surely in the norm of an L_2 space. Two general results, on which the development of this paper is based, are also presented.

INTRODUCTION

Series representations of stochastic processes are of considerable use in several problems of statistical communication theory. If $\{x(t,\omega), t\epsilon T\}$ is a stochastic process of second order defined on the Lebesgue measurable set T of the real line, a general series representation is given by

$$x(t,\omega) = \sum_k a_k(t)\, \eta_k(\omega) \tag{1}$$

where the convergence is usually in the stochastic mean. The significance of such a representation is in the fact that it decomposes the stochastic process into the random, time independent, part $\{\eta_k(\omega)\}$ and the time dependent deterministic part $\{a_k(t)\}$, thus providing some insight into the structure of the stochastic process as well as a tool for the study of many particular problems.

Various series representations have been obtained under certain assumptions. Before mentioning them, let us introduce some notation. Let R^1 be the real line, $\mathcal{B}^1$ the σ-algebra of Lebesgue measurable sets on R^1, $\mathcal{B}^1{}_T$ the restriction of $\mathcal{B}^1$ to $T\epsilon\mathcal{B}^1$, R^2 the plane, $\mathcal{B}^2$ the σ-algebra of Lebesgue sets on R^2 and m the Lebesgue measure on the real line.

A second order, mean square continuous stochastic process $\{x(t,\omega), t\epsilon I\}$ on a compact interval of the real line, has the series representation [3,5]

$$x(t,\omega) = \sum_k f_k(t)\, \xi_k(\omega) \tag{2}$$

where the convergence is in the stochastic mean uniformly on I. $\{f_k(t)\}_k$ and $\{\lambda_k\}_k$ are the corresponding eigenfunctions and nonzero eigenvalues of the operator on $L_2(I, \mathcal{B}^1{}_I, m)$ with kernel $R(t,s)$, where $R(t,s)$ is the autocorrelation function of $x(t,\omega)$. $\{\xi_k(\omega)\}_k$ is a set of random variables defined by

$$\xi_k(\omega) = \int_I x(t,\omega)\, f_k^*(t)\, m(dt) \tag{3}$$

---

[*] This author's work was partly supported by the National Science Foundation under Grant GU-2059.
[**] This author's work was supported by the National Science Foundation under Grant GK-13192.

in the stochastic mean and satisfying $E[\xi_k \xi^*_\ell] = \lambda_k \delta_{k\ell}$. Representation (2) is known as the Karhunen-Loève expansion and its great advantage is that both the time functions and the random variables are orthogonal.

The Karhunen-Loève representation makes the minimum of assumptions on the process, but in general it is not valid over the entire real line or over non-compact intervals of the real line. A series representation similar to (2) and valid over a possibly infinite interval $I = (a,b)$, $-\infty \le a < b \le +\infty$, of the real line is given in [3] for the subclass of second order, mean square continuous stochastic processes $\{x(t,\omega), t\epsilon I\}$ whose autocorrelation functions satisfy

$$\int_I R(t,t)\, m(dt) < \infty. \tag{4}$$

Equations (2) and (3) are valid in this case, but the convergence in (2) is not in general uniform on I. If I is a compact interval, then (4) is satisfied and the convergence in (2) is uniform.

However, for any Lebesgue measurable set T of the real line, it is known that a second order stochastic process $\{x(t,\omega), t\epsilon T\}$ has a series representation of the form (1) if and only if the Hilbert space $H(x,T)$ spanned in the mean square sense by the random variables $\{x(t,\omega), t\epsilon T\}$ is separable [4,p.27]. On the other hand, if $\{x(t,\omega), t\epsilon T\}$ is weakly continuous on T, then $H(x,T)$ is separable [7,Thm.2D]. Hence every weakly continuous stochastic process, and a fortiori every mean square continuous stochastic process, has a series representation of the form (1) on any Lebesgue set T. What remains to be established then, is an explicit way of determining the time functions $\{a_k(t)\}$ and the random variables $\{\eta_k(\omega)\}$. Let us note that if the set $\{\eta_k(\omega)\}$ is orthonormal, then the time functions are given by

$$a_k(t) = E[x(t)\, \eta_k{}^*]. \tag{5}$$

The time functions and the random variables in the representation (1), valid over the entire real line, have been obtained in [6] for the class of mean square continuous, wide sense stationary stochastic processes, and in [2] for the class of harmonizable stochastic processes.

A general way to obtain complete sets of random variables in the span $H(x,T)$ of any weakly continuous stochastic process $\{x(t,\omega), t\epsilon T\}$, and therefore series representations of the form (1), is presented in Theorem 4. The random variables are defined by linear operations on the stochastic process. The significance of this novel approach in obtaining series representations for weakly continuous stochastic processes is illustrated by the representations derived in subsequent theorems as immediate consequences of Theorem 4. Theorems 5 and 6 provide two series representations for any weakly continuous stochastic process; that is, they provide two distinct ways of determining the time functions and the random variables in (1). The time functions in the representation of Theorem 5 can be obtained explicitly in a straightforward way, while in order to obtain the time functions in the representation of Theorem 6 the computation of the eigenfunctions of an integral equation is required. The novelty of the representations presented in Theorems 5 and 6 is in the fact that they hold for all weakly, and therefore mean square, continuous stochastic processes over a broad class of Lebesgue sets of the real line, finite or infinite, compact or non-compact. In contrast, the representations of mean square continuous processes given in [3,5] hold either only on compact subsets of the real line or on any Lebesgue set of the real line with the additional condition (4) and they are included in Theorem 6 as particular cases. Both representations given in Theorems 5 and 6 can be written in the form of the representations given in [6] for mean square continuous, wide sense stationary stochastic processes, and in [2] for harmonizable stochastic processes. They also provide two absolutely convergent series representations for $R(t,s)$.

For all measurable second order stochastic processes, a series representation of the form (1) is obtained in Section 4, where the convergence is not in the stochastic mean anymore but in the norm of an appropriate L_2 space almost surely. It is also shown that both representations of Theorems 5 and 6 converge also almost surely in an L_2 space sense.

All the results presented in Sections 3 and 4 rely on two general results presented as Theorems 1 and 2, in Section 2. They essentially say that for every measurable second order stochastic process defined on any Lebesgue measurable set T of the real line, there exists an appropriate L_2 space over T with the properties that (i) almost all sample functions of the process belong to it, and (ii) the autocorrelation function of the process, considered as kernel, defines a completely continuous operator on it of the trace class. It is believed that the significance of these two results is beyond their implications presented in this paper.

2. SOME GENERAL RESULTS

This section presents two general results on measurable, second order stochastic processes defined on any Lebesgue set T of the real line, that are used in subsequent sections.

THEOREM 1. For every measurable, second order stochastic process $\{x(t,\omega),\ t\epsilon T\}$ there exists a finite, non-negative measure μ on $(T, \mathcal{B}^1{}_T)$ such that

$$\int_T R(t,t)\ \mu(dt) < \infty \tag{6}$$

where $R(t,s)$ is the autocorrelation function of $x(t,\omega)$.

Proof: Define a function f on T by

$$f(t) = \begin{cases} 1 & \text{for } 0 \le R(t,t) \le 1 \\ 1/R(t,t) & \text{for } R(t,t) > 1. \end{cases} \tag{7}$$

Then f is measurable and $0 \le R(t,t)\ f(t) \le 1$ on T. If ν is any finite, non-negative measure on $(T,\mathcal{B}^1{}_T)$ and if the measure μ on $(T,\mathcal{B}^1{}_T)$ is defined by $[d\mu/d\nu] = f$, then (6) is satisfied. Furthermore, μ is non negative and finite since $0 \le f(t) \le 1$ on T. Q.E.D.

Let us point out that if $R(t,s)$ is uniformly bounded on $T\times T$, then μ can be chosen to be any finite, non-negative measure on $(T,\mathcal{B}^1{}_T)$. This includes the cases where x is mean square continuous, wide sense stationary or harmonizable. If $R(t,t)$ is integrable over T, then μ can be chosen to be the Lebesgue measure m on T.

If the measure ν is chosen to be absolutely continuous with respect to the Lebesgue measure m with Radon-Nikodym derivative ϕ, then μ will have Radon-Nikodym derivative with respect to m:

$$[d\mu/dm](t) = F(t) = f(t)\ \phi(t). \tag{8}$$

$f(t)$ may be chosen positive a.e. $[m]$, as for instance in (7). If ϕ is also chosen non-zero on T a.e. $[m]$, then the function F defined on T by (8) will be non-zero a.e. $[m]$, non-negative and Lebesgue integrable over T. Hence we have the following

COROLLARY. The measure μ of Theorem 1 can always be chosen to be absolutely continuous with respect to the Lebesgue measure m such that $[d\mu/dm](t) = F(t) \neq 0$ a.e. $[m]$ on T.

It should be pointed out that the measure μ of Theorem 1 is not uniquely determined. In the construction of μ given in the proof of Theorem 1, the choice of f is clearly not unique and ν is an arbitrary finite measure; neither is this construction the only way to obtain a measure μ with the properties stated in Theorem 1. However, it should be emphasized that the considerable freedom in the choice of the measure μ with the properties stated in Theorem 1, far from being a drawback, represents an advantage to be appropriately exploited in every particular case.

THEOREM 2. Let $\{x(t,\omega),\ t\in T\}$ be a measurable, second order stochastic process and let μ be the measure introduced in Theorem 1. Then $x(t,\omega) \in L_2(T,\mathcal{B}^1{}_T,\mu) = L_2(\mu)$ almost surely.

Proof: The measurability of x and Tonelli's theorem imply

$$\int_T R(t,t)\mu(dt) = \int_T E[|x(t,\omega)|^2]\mu(dt) = E[\int_T |x(t,\omega)|^2 \mu(dt)]. \tag{9}$$

Hence, by (9), $x(t,\omega) \in L_2(\mu)$ a.s. Q.E.D.

It should be pointed out that the applicability of the property proven in Theorem 2 is far beyond its use in Section 4. For instance, this property can be used in generalizing a number of results established in the literature in the particular case of stochastic processes whose sample functions are almost surely square integrable with respect to the Lebesgue measure.

3. DERIVATION OF SERIES REPRESENTATIONS

Throughout this and the next section, the following notation is used:

(i) The set T is any Lebesgue measurable set of the real line such that no subset of T open relative to T has zero Lebesgue measure. In particular, T may be an interval, finite or infinite, compact or non-compact. Notice that this condition is not needed in Theorems 3 and 7.

(ii) $\{x(t,\omega),\ t\in T\}$ is a measurable, second order stochastic process defined on the Lebesgue measurable sets T of assumption (i).

(iii) μ is a non-negative measure on $(T,\mathcal{B}^1{}_T)$ such that (6) is satisfied and $[d\mu/dm](t) \neq 0$ a.e. $[m]$ on T. As shown in Section 2 (Theorem 1 and its corollary), such a measure always exists and can be found explicitly if $R(t,t)$ is known. Notice that in order to include in the subsequent study the case considered in [3], i.e., $\mu = m$, the assumption is not made that μ is finite.

Our goal, as made clear in the introduction, is to find complete sets of random variables in the span $H(x,T)$ of a second order stochastic process. A convenient set of random variables $\{\eta_k(\omega)\}_k$ is defined by

$$\eta_k(\omega) = \int_T x(t,\omega)\, \phi_k{}^*(t)\, \mu(dt) \tag{10}$$

almost surely, where $\{\phi_k(t)\}_k$ is an arbitrary complete set of functions in $L_2(\mu)$.

THEOREM 3. The random variables $\{\eta_k(\omega)\}_k$ defined by (10) are of second order. If $H(\eta)$ is the Hilbert space spanned in the mean square sense by the random variables $\{\eta_k(\omega)\}_k$, then

$$H(\eta) \subseteq H(x,T). \tag{11}$$

Proof: Eq. (6) implies $R(t,s) \in L_2(T\times T, \mathcal{B}^2{}_{T\times T}, \mu\times\mu) = L_2(\mu\times\mu)$. Hence $\int_T\int_T R(t,s)\phi_k^*(t)\phi_k(s)\mu(dt)\mu(ds)$ is finite and the random variables $\{\eta_k(\omega)\}_k$ are of second order. To show (11), it suffices to show that $\eta_k \in H(x,T)$ for all k, or that if η_k is orthogonal to $x(t)$ for all $t\in T$ then $E[|\eta_k|^2] = 0$. Assume that η_k is orthogonal to $x(t)$ for all $T\in T$. Then

$$0 = E[x(t)\eta_k^*] = \int_T R(t,s)\phi_k(s)\mu(ds) = (R\phi_k)(t) \tag{12}$$

for all $t\in T$, and

$$\begin{aligned} E[|\eta_k|^2] &= \int_T \int_T R(t,s)\ \phi_k^*(t)\ \phi_k(s)\ \mu(dt)\ \mu(ds) \\ &= (R\phi_k,\ \phi_k)_{L_2(\mu)} = 0. \qquad \text{Q.E.D.} \end{aligned} \tag{13}$$

A natural question to ask is under what conditions equality holds in (11). Before pursuing this question further let us make the following remark. If $T'\subset T$ is such that $m(T\sim T') = 0$, then $H(\eta) \subseteq H(x,T')$. This is shown as Theorem 3; the only difference is that in this case $(R\phi_k)(t) = 0$ for $t\in T'$, i.e., a.e. [m] on T, but this suffices to imply $R\phi_k = 0$ in $L_2(\mu)$ and hence $E[|\eta_k|^2] = 0$. This remark suggests that $H(\eta)$ may not contain the random variables $x(t)$ corresponding to points of discontinuity of x of some kind, for instance discontinuity in the mean square sense; it also suggests that $H(\eta)$ contains only the "smooth" part of x and therefore in order to have $H(\eta) = H(x,T)$ the process x has to be "smooth" in some sense. A continuity ("smoothness") condition on x which implies $H(x,T) = H(\eta)$ is given in the next theorem. Let us note that a stochastic process $\{x(t,\omega),\ t\in T\}$ is called weakly continuous on T if for every $t\in T$ and every $\xi \in H(x,T)$, $\lim_{T\ni\tau\to t} E[x(\tau)\xi^*] = E[x(t)\xi^*]$.

THEOREM 4. If $\{x(t,\omega),\ t\in T\}$ is weakly continuous on T, then

$$H(\eta) = H(x,T). \tag{14}$$

Proof: In view of (11), it suffices to show $H(x,T) \subseteq H(\eta)$, or equivalently that if $\zeta \in H(x,T)$ is orthogonal to η_k for all k then $E[|\zeta|^2] = 0$. Assume that $\zeta \in H(x,T)$ is orthogonal to η_k for all k. Then

$$0 = E[\zeta\eta_k^*] = \int_T E[\zeta x^*(t)]\ \phi_k(t)\ \mu(dt) \tag{15}$$

for all k. The function $f(t) = E[x(t)\zeta^*]$ belongs to $L_2(\mu)$ since

$$\int_T |f(t)|^2\ \mu(dt) \le E[|\zeta|^2] \int_T R(t,t)\ \mu(dt) < +\infty \tag{16}$$

and is orthogonal to the complete set $\{\phi_k(t)\}_k$ by Eq. (15). Hence $|f(t)|^2 = 0$ on T a.e. [μ], or by (8) $|f(t)|^2\ F(t) = 0$ on T a.e. [m], and since $F(t) \neq 0$ on T a.e. [m], it follows that $f(t) = 0$ on T a.e. [m]. We further note that $f(t)$ is a continuous function since x

is weakly continuous. It follows from this and assumption (i) that $f(t) = 0$ for all $t \in T$ so that $\zeta(\omega)$ is orthogonal to the set $\{x(t,\omega),\ t \in T\}$, i.e., to $H(x,T)$. Thus $E[|\zeta|^2] = 0$. Q.E.D.

It should be pointed out that Theorem 4 and all subsequent results of this section, are true a fortiori for mean square continuous processes, since mean square continuity implies weak continuity. It would be interesting to find a necessary and sufficient condition for $H(x,T) = H(\eta)$. The weak continuity of x is shown here to be only a sufficient condition.

Theorem 4 implies that every weakly continuous process x admits the representation

$$x(t,\omega) = \sum_k a_k(t)\, \eta_k(\omega); \quad t \in T \tag{17}$$

in the stochastic mean on T. However, the time functions $\{a_k(t)\}_k$ in (17) cannot be explicitly obtained because the set $\{\eta_k(\omega)\}_k$ is not orthogonal. An explicit expression for the time functions and the random variables in the representation (1) is given in Theorem 5, which follows directly from Theorem 4.

THEOREM 5. Every stochastic process $\{x(t,\omega),\ t \in T\}$ weakly continuous on T admits the representation

$$x(t,\omega) = \sum_k b_k(t)\, \xi_k(\omega); \quad t \in T \tag{18}$$

where the convergence is in the stochastic mean on T. The random variables $\{\xi_k(\omega)\}_k$ are derived from $\{\eta_k(\omega)\}$ by the Gram-Schmidt orthonormalization procedure: $\xi_k(\omega) = \sum_{\ell=1}^{k} c_{k\ell}\eta_\ell(\omega)$. They are given by

$$\xi_k(\omega) = \int_T x(t,\omega)\, g_k{}^*(t)\, \mu(dt) \tag{19}$$

almost surely, where $g_k(t) = \sum_{\ell=1}^{k} c_{k\ell}{}^*\phi_\ell(t)$, satisfy $E[\xi_k\xi_\ell{}^*] = \delta_{k\ell}$ and constitute a complete set in $H(x,T)$. The time functions $\{b_k(t)\}_k$ are given by

$$b_k(t) = \int_T R(t,s)\, g_k(s)\, \mu(ds). \tag{20}$$

Also $R(t,s)$ admits the representation

$$R(t,s) = \sum_k b_k(t)\, b_k{}^*(s); \quad t,s \in T \tag{21}$$

where the convergence is absolute in t and s on $T \times T$.

In the case where μ is a finite measure, as a complete set of functions in $L_2(T,\mathcal{B}^1{}_T,\mu)$ one can use the general orthonormal and complete set of functions $\{\phi_k(t)\}_k$ given in [6,Thm.2] by

$$\{\phi_k(t) = \mu^{-\frac{1}{2}}(T)\ exp[ik2\pi\mu\{(-\infty,t) \cap T\} / \mu(T)]; \quad k = 0,\pm1,\pm2,\ldots\}. \tag{22}$$

If μ is the Lebesgue measure and T is the entire real line or the half line then as a complete set $\{\phi_k(t)\}_k$ one can choose the Tchebysheff-Hermite or the Tchebysheff-Laguerre functions respectively. On the other hand, if T is a finite interval then the Legendre polynomials or the trigonometric functions are applicable.

Upon making a particular choice for the complete set of functions $\{\phi_k(t)\}_k$ used in (10), a representation very similar to the usual

Karhunen-Loève representation is obtained in Theorem 6 and is valid over the class of Lebesgue measurable sets of the real line defined by assumption (i). The results are similar to those in [3]. However, the representation derived in Theorem 6 applies to all weakly continuous stochastic processes, all mean square continuous processes included, while the representation of [3] applies to the class of mean square continuous stochastic processes satisfying (4), which clearly does not contain the entire class of mean square continuous stochastic processes.

THEOREM 6. Every stochastic process $\{x(t,\omega),\ t \in T\}$ weakly continuous on T admits the representation

$$x(t,\omega) = \sum_k f_k(t)\ \xi_k(\omega); \quad t \in T \tag{23}$$

where the convergence is in the stochastic mean on T. $\{f_k(t)\}_k$ and $\{\lambda_k\}_k$ are the corresponding eigenfunctions and non-zero eigenvalues of the operator on $L_2(\mu)$ with kernel $R(t,s)$. The random variables $\{\xi_k(\omega)\}_k$ are defined by

$$\xi_k(\omega) = \int_T x(t,\omega)\ f_k^*(t)\ \mu(dt) \tag{24}$$

almost surely, satisfy $E[\xi_k \xi_\ell^*] = \lambda_k \delta_{k\ell}$ and constitute a complete set in $H(x,T)$. Also $R(t,s)$ admits the representation

$$R(t,s) = \sum_k \lambda_k\ f_k(t)\ f_k^*(s); \quad t,s \in T \tag{25}$$

where the convergence is absolute in t and s on $T \times T$.

Proof: It follows from $R(t,s) \in L_2(\mu\times\mu)$ that $R(t,s)$ is the kernel of a Hilbert-Schmidt operator R from $L_2(\mu)$ to $L_2(\mu)$. Hence R is a self-adjoint, non-negative definite, completely continuous operator [1,pp. 54-56, 58-59]; it is also of the trace class. Its non-zero eigenvalues $\{\lambda_k\}_k$ and the corresponding eigenfunctions $\{f_k(t)\}_k$ satisfy the integral equation $Rf_k = \lambda_k f_k$ [1,pp.124-129]. The set $\{f_k(t)\}_k$ is orthonormal in $L_2(\mu)$ and complete in the range of R. Let $\{h_\ell(t)\}_\ell$ be an orthonormal basis in the orthogonal complement of the range of R in $L_2(\mu)$. By applying Theorem 4, we obtain

$$x(t,\omega) = \sum_k a_k(t)\ \xi_k(\omega) + \sum_\ell d_\ell(t)\ \zeta_\ell(\omega); \quad t \in T \tag{26}$$

in the stochastic mean, where the ξ_k's are given by (24) and $\zeta_\ell(\omega) = \int_T x(t,\omega) h_\ell^*(t) \mu(dt)$ almost surely. Since $Rh_\ell = 0$ in $L_2(\mu)$ for all ℓ, it follows that $E[|\zeta_\ell|^2] = (Rh_\ell, h_\ell)_{L_2(\mu)} = 0$ for all ℓ. Also $\lambda_k a_k(t) = E[x(t)\xi_k^*] = (Rf_k)(t) = \lambda_k f_k(t)$ for all k and $t \in T$. Hence (23) follows from (26). Q.E.D.

Remark 1: In comparing the two representations given in Theorems 5 and 6, the following should be noticed: (i) In both representations, the random variables are orthogonal. (ii) The time functions in the representation of Theorem 6 are orthogonal, while in the representation of Theorem 5 they are not. (iii) The representation of Theorem 6 requires the computation of the eigenfunctions of an integral equation, while the representation of Theorem 5 can be obtained explicitly in a straightforward way.

Remark 2: For series of independent random variables, it is known [5, p.251] that convergence in the mean square sense implies almost sure convergence. It follows that, if the weakly continuous stochastic process x

is Gaussian, then both representations (18) and (23) of Theorems 5 and 6 converge almost surely for every $t \in T$.

Remark 3: By using the properties of the autocorrelation function of a weakly continuous stochastic process [7,Thm.2E], it can be shown that every function $f(t)$ of the form $f = Rg$, where $g \in L_2(\mu)$, is continuous in t on T. In particular, the functions $b_k(t)$ in the representation (18) of Theorem 5, given by (20), as well as the eigenfunctions $f_k(t)$ of $R(t,s)$ used in the representation (23) of Theorem 6, are continuous functions of t on T.

Remark 4: If x is mean square continuous on T, it can be shown, by using Dini's theorem [8,p.136], that the convergence in both representations of x given in Theorems 5 and 6 is uniform on compact subsets of T. The same is true for the representations of $R(t,s)$, (21) and (25). If in particular T is compact, then an alternative proof to the Karhunen-Loève representation is thus obtained, a proof which does not rely upon Mercer's theorem. Moreover, an alternative proof to Mercer's theorem is obtained.

Remark 5: As a brief comment on the relationship of the main earlier representations [3,5,6,2] to those derived in Theorems 5 and 6 the following should be noted. Karhunen's representation [3] of mean square continuous stochastic processes satisfying (4) is clearly a particular case of Theorem 6. As noted in Remark 4, the so-called Karhunen-Loève representation is also a particular case of Theorem 6. Also, it is easily seen that both representations of Theorems 5 and 6 can be written in the form of the representations derived in [6] and in [2] over the entire real line for the class of mean square continuous, wide sense stationary stochastic processes and for the class of harmonizable stochastic processes respectively.

4. SAMPLE FUNCTION REPRESENTATION

In this section, the process x, the set T and the measure μ are defined as in the beginning of Section 3. The fact that almost all sample functions of the process x belong to $L_2(\mu)$, shown in Theorem 2 of Section 2, implies in a straightforward way the following.

THEOREM 7. If $\{\phi_k(t)\}_k$ is an arbitrary complete set of orthonormal functions in $L_2(\mu)$, then x admits the representation

$$x(t,\omega) = \sum_k \phi_k(t)\, \eta_k(\omega) \tag{27}$$

where the convergence is in $L_2(\mu)$ almost surely, and the random variables $\{\eta_k(\omega)\}_k$ are given almost surely by

$$\eta_k(\omega) = \int_T x(t,\omega)\, \phi_k{}^*(t)\, \mu(dt). \tag{28}$$

The remark which follows Theorem 5 on how the functions $\{\phi_k(t)\}_k$ can be found in all cases under consideration, applies here also. The representation of Theorem 7 clearly applies to weakly continuous processes x. For this class of processes, it is now shown that both representations given in Theorems 5 and 6 are also sample function representations.

THEOREM 8. For every weakly continuous stochastic process x the representations (18) and (23) given in Theorems 5 and 6 converge also in $L_2(\mu)$ almost surely.

Proof: (i) For the representation (18) of Theorem 5. Consider the functions $\{b_k(t)\}_k$ given by (20): $b_k = Rg_k$, and let $\{d_\ell(t)\}_\ell$ be an orthonormal basis in the orthogonal complement of the span of the set $\{b_k(t)\}_k$ in $L_2(\mu)$. Then we have

$$x(t,\omega) = \sum_k b_k(t)\, \beta_k(\omega) + \sum_\ell d_\ell(t)\, \delta_\ell(\omega) \tag{29}$$

in $L_2(\mu)$ almost surely (a.s.). Eq. (20) implies that $\beta_k(\omega) = \xi_k(\omega)$ a.s., where the ξ_k's are given by (19). Also $\delta_\ell(\omega) = \int_T x(t,\omega) d_\ell^*(t)\mu(dt)$ a.s. It follows that $E[\delta_\ell \xi_k^*] = (Rg_k, d_\ell)_{L_2(\mu)} = (b_k, d_\ell)_{L_2(\mu)} = 0$ for all k and ℓ, and hence δ_ℓ is orthogonal to $H(x,T)$ for all ℓ. Since Theorem 3 implies $\delta_\ell \in H(x,T)$ for all ℓ, we have $E[|\delta_\ell|^2] = 0$ for all ℓ. It follows now from (29) that (18) converges in $L_2(\mu)$ a.s.

(ii) For the representation (23) of Theorem 6. Apply Theorem 7 for the complete orthonormal set $\{f_k(t)\}_k \cup \{h_\ell(t)\}_\ell$ in $L_2(\mu)$ introduced in the proof of Theorem 6 and proceed as in (i). Q.E.D.

REFERENCES

1. N.I. Akhiezer and I.M. Glazman, Theory of Linear Operators in Hilbert Space, Vol. 1, Ungar, New York, 1961.

2. S. Cambanis and B. Liu, On Harmonizable Stochastic Processes, Inform. and Control, Vol. 17, No. 2, September 1970.

3. K. Karhunen, Zur Spectraltheorie stochastischer Prozesse, Ann. Acad. Scient. Fennicae, Ser. AI, No. 34, (1946), pp. 1-7.

4. K. Karhunen, Über lineare Methoden in der Wahrscheinlichkeitsrechnung, Ann. Acad. Scient. Fennicae, Ser. AI, No. 37, (1947), pp. 1-79.

5. M. Loève, Probability Theory, Van Nostrand, Princeton, New Jersey, 1963.

6. E. Masry, B. Liu and K. Steiglitz, Series Expansion of Wide-Sense Stationary Random Processes, IEEE Trans. Information Theory, Vol. IT-14 (1968), pp. 792-796.

7. E. Parzen, Statistical Inference on Time Series by Hilbert Space Methods, in Time Series Analysis Papers, E. Parzen, Holden-Day, San Francisco, 1967, pp. 251-382.

8. W. Rudin, Principles of Mathematical Analysis, McGraw-Hill, New York, 1964.

COMPUTER AIDED DESIGN WITH A MINI-COMPUTER

TOWNSEND D. MAC COUN
Quindar Electronics, Inc.
Springfield, New Jersey 07081

ABSTRACT

Three years experience in using a mini-computer in a moderate size R & D laboratory, for the design of filters, networks and other projects have been economical, reliable and productive.

INTRODUCTION

The choice of a computer, is of course, influenced to a large extent by the nature of the company's business. Quindar Electronics are producers of industrial control, telemetering and supervisory control equipment. We are also producers of telephone equipment, data modems and other communication products.

We therefore have large requirements for filters, equalizers, oscillators and discriminators for use in our own equipment. Some of the larger systems require several hundred tone channels.

In some of the larger supervisory systems, a mini-computer is an intergral component for logging production data, and various kinds of alarms.

Ease of interfacing with the real world, for gathering data was, therefore, of considerable influence. Other factors were, the software languages provided, the software support, and the reputation for reliability of the manufacturer.

The computer which was chosen for this purpose was the Hewlett-Packard 2115 with 8K of 16 bit memory. The languages are Basic, Fortran, (an extended version of Fortran II,) Algol and their assembly language.

It is highly desirable to have a high speed paper tape reader so that programs can be loaded in approximately one minute.

FILTERS

The computer has principally been used for the design of filters for production and use in our equipment. These filters generally are either third order or fourth order Chebycheff or fourth order eliptical filters operating in the audio range. We have written, approximately, 75 different filter programs of which, at the present time, there are 19 currently in very active use. The programs are written so that together with the proper circuit diagram they present all of the information necessary for the production department to build and test the filters.

In the format of the print-out the final line is a form number. The print-out is pasted to a copy of this form. It contains the circuit, the arrangement of the components and a response curve with all the tolerances included. A xerox copy is then made and issued to the production department. This avoids processing through the drafting department, and the possibility of errors in transferring the data to a drawing. It also reduces the processing time to about 15 minutes.

Figure 1 shows a typical control print. A preferred list of 63 capacitor values are stocked and marked by the manufacturer with a part number and a dash number. An eleven word subroutine with an algorithm to select the nearest dash number is used (as an option) to print the dash number for all capacitors above 1000 pf and the capacity value below 1000 pf.

There are 10 series of receiver bandpass filter with bandwidths from 50 to 1360 Hz. Four of these series have two different slopes and stop bands - a total of 370 filters - for each filter there are corresponding transmitter filters, oscillator networks and discriminators.

One of the problems encountered early in this work was the bridging of the gap between the truly arithmetical design of the filter and the judgment that has to be used in introducing transformations such as Bingham's transformation, Norton's transformation and various frequency scaling operations. In order to accomplish this, we had to sit down and draw flow charts of what our intuitive thinking process was in arriving at the right transformation ratios. This was very instructive and challenging. These programs are now completely automated so that they introduce transformations that always give realizable components.

Most of the programs have options to print out the 600 ohm or normalized values and the option to print capacitor values or dash numbers. Low pass filters have the option for minimum capacitor configuration and the high pass filters have the option for star-delta transformations. Several programs have the option for the zig-zag transformation.

ANALYSIS PROGRAMS

It was recognized early in this work, that we had an urgent need for a reliable circuit analysis program. The size of our computor memory precluded the use of the more elegant programs developed by Calahan (1), Pottle (2) and others. We therefore settled for a program that is restricted to a ladder of up to 10 branches with any one of eight circuits in each branch. The algorithm is very similar to that developed by Bashkow (3), alternately calculating admittances and impedances. This covers almost all of the pracitcal filter designs and by an extension of segmenting the program, we can extend this on up to 20 branches. The input is a list of the element values, either real or normalized, the input and output impedance and input frequency. Four options are provided: 1 - R in ohms or one value of Q for all inductors and a second value of Q for all capacitors. 2 - Element values real or normalized. 3 - Frequency incremented starting at f or read from a selected list on a data tape. 4 - Envelope delay required or not.

The output of this program is the attenuation in db, the impedance looking into the filter, (real, imaginery and the magnitude,) and the envelope delay. When proper values are chosen, and the filter is built, the agreement between the measured response and the analysis has been within a db at all points. The option to enter Q values has added immeasurably to the usefulness of this program.

A program to analyze a network using the pole, zero locations has also been developed.

A program to analyze and optimize phase locked oscillator circuits has been most successful. It converges on a set of non-linear equations.

Several programs for the design and evaluation of active filters are in current use.

RELIABILITY STUDIES

To establish a reliability criterion for our supervisory and transfer trip systems, two different programs have been used.

The coding structure of a supervisory system typically consists of a sync bit, a station address, a point address, parity bits and a command. The program exercises the equipment by generating a succession of codes, transmitting them over our tone channel and then compares the received code for errors. The telephone line is simulated by attenuators, and a white noise generator is bridged across the line so that the signal to noise ratio can be controlled. Using this program over a weekend, showed that approximately 3,800,000 messages had been sent. 1003 errors in transmission occurred and only 6 had gotten past the parity and other checks for validity. At each failure the transmitted and received bit structure were printed out. This test was repeated at various noise levels. Analysis of the nature of the failures helped to pinpoint the cause of the failures. This resulted in minor circuit changes which substantially increased system reliability.

The second program was a mathematical-statistical model of the transfer trip system. One of the pitfalls of this kind of method, the assumption that the noise output of a filter is gaussian, was avoided. The program provided for the insertion of the transfer function of the filters and the ability to alter certain parameters. A white noise source was simulated by using a random number generator, ahead of the transfer function equations.

The results of running this program were in close agreement with the results of the first program and were indicative of a valid model. Manipulation of the parameters again resulted in improved system performance and a realistic criterion for evaluating various system performances.

TRANSLATOR & EDIT PROGRAMS

Programs have been written in assembly language for translation between languages, for instance, between the code used on IBM's Selectric typewriters and standard ASCII code.

An Edit program for searching a program for a six character string and replacing it with another string is in preparation.

A program to resequence a program in Basic and correctly replace all reference line numbers with new numbers has been developed. It has been given to the Hewlett-Packard users library and has had a considerable distribution.

PRODUCTION TESTING

The production department also has made considerable use of the computer and we expect it will make considerably more in the future. Programs have been written in a special language for which we have written our own compiler for testing printed circuit boards. This special language is oriented to the logic designers and testers use. It consists of only six instructions.

They are:

CLEAR	This sets all pin numbers high
SET	as SET P2-6 LO
WAIT (100)	This waits 100 ms or anything up to 3000 ms
TEST	as TEST P1-C HI
WHEN	as WHEN P2-N LO to wait for a mono flip flop to change
PAUSE	This waits for a new RUN signal

Tests can be made on a printed circuit board where approximately 500 tests on a board can be completed in approximately one second in the event that the board is good. If the board has failures in it, it will print out which tests failed. The board to be tested is plugged into a test jig which is interfaced to the computer. If the board tests OK, the teletype bell rings.

ALTERNATE POSSIBILITIES

As time has gone on and the requirements for networks and for production uses have become more sophisticated, we will have to go to, either a larger computer or to the addition of some time sharing services.

The large analysis programs, covering circuits that are not ladder networks, and which allow the inclusion of diodes and active elements, as well as the optimization programs offered by the time sharing systems are very attractive.

At the present time we are evaluating these possibilities.

CONCLUSION

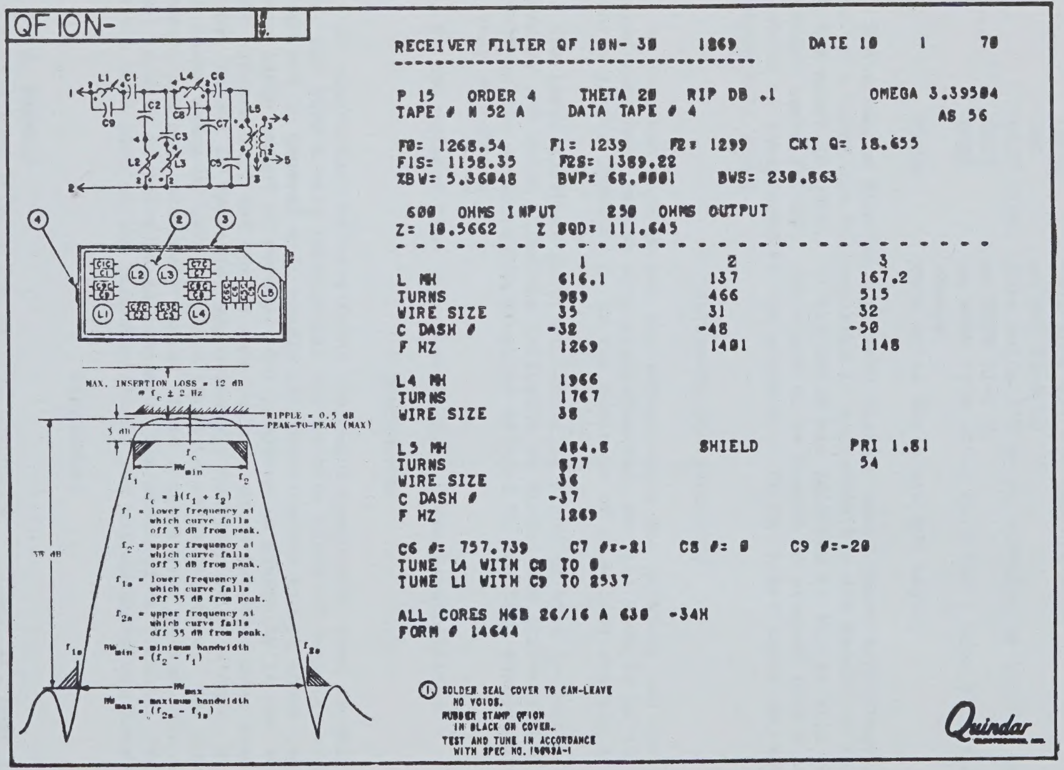
In conclusion, we have found that mini-computers, even of this small size, can have a very substantial use in this field of computer-aided design and in general will handle all requirements except those requiring a very large amount of memory. Our computer is generally in use seven or eight hours a day, and during some periods which extended over some six months, it was in operation approximately fourteen hours a day. With all this heavy usage we have worn out three teletype machines. The down time, however, has been extremely small. Over a three year period, the down time, because of teletype maintenance, has amounted to perhaps, four days and the down time on the computer itself has not exceeded six hours.

REFERENCES

3. T. R. BASHKOW
Mathematical Methods for Digital Computers Chap. 26, page 280
Ralston and Wilf John Wiley & Sons, Inc., N.Y.C.

1. DONALD A. CALAHAN
Modern Network Synthesis Hayden Book Co., Inc. , N.Y.C.

2. R. C. ALLEN and C. POTTLE
Stable Integration Methods for Electronic Circuit Analysis with Widely Separated Time Constance Page 311
Sixth Annual Allerton Conference


QF 10N-
L1
C1
L4
C6
C2
C8
C7
C9
C3
L5
L2
L3
C5
1
2
3
4
5
④
②
③
L1
L2
L3
L4
L5
MAX. INSERTION LOSS = 12 dB
@ f_c ± 2 Hz
RIPPLE = 0.5 dB
PEAK-TO-PEAK (MAX)
3 dB
35 dB
f_c
f_1
f_2
BW_{min}
f_{1s}
f_{2s}
BW_{max}

$f_c = \frac{1}{2}(f_1 + f_2)$

f_1 = lower frequency at which curve falls off 3 dB from peak.

f_2 = upper frequency at which curve falls off 3 dB from peak.

f_{1s} = lower frequency at which curve falls off 35 dB from peak.

f_{2s} = upper frequency at which curve falls off 35 dB from peak.

BW_{min} = minimum bandwidth $= (f_2 - f_1)$

BW_{max} = maximum bandwidth $= (f_{2s} - f_{1s})$

```
RECEIVER FILTER QF 10N- 30    1269        DATE 10    1     70
------------------------------------

P 15    ORDER 4    THETA 20   RIP DB .1          OMEGA 3.39504
TAPE # N 52 A     DATA TAPE # 4                          AS 56

F0= 1268.54     F1= 1239    F2= 1299    CKT Q= 18.655
FIS= 1158.35     F2S= 1389.22
%BW= 5.36048     BWP= 68.0001    BWS= 230.863

 600  OHMS INPUT     250  OHMS OUTPUT
Z= 10.5662    Z SQD= 111.645
- - - - - - - - - - - - - - - - - - - - - - - - - - - - -
                      1                2                3
L MH                616.1            137              167.2
TURNS               989              466              515
WIRE SIZE           35               31               32
C DASH #           -32              -48              -50
F HZ                1269             1401             1148

L4 MH               1966
TURNS               1767
WIRE SIZE           38

L5 MH               484.8          SHIELD           PRI 1.81
TURNS               877                              54
WIRE SIZE           36
C DASH #           -37
F HZ                1269

C6 #= 757.739    C7 #=-21    C8 #: 0    C9 #=-20
TUNE L4 WITH C8 TO 0
TUNE L1 WITH C9 TO 2537

ALL CORES M6B 26/16 A 630  -34H
FORM # 14644
```

① SOLDER SEAL COVER TO CAN-LEAVE NO VOIDS.
RUBBER STAMP QF10N IN BLACK ON COVER.
TEST AND TUNE IN ACCORDANCE WITH SPEC NO. 14643A-1

Quindar

FORM 14644A/A

Fig. 1

THE OPTIMIZATION OF NUMERICAL MODELS OF SEMICONDUCTOR DEVICES*

L. C. McAFEE, JR.
Department of Electrical Engineering
The University of Michigan
Ann Arbor, Michigan

ABSTRACT

For semiconductor devices analytic solutions to the set of describing nonlinear partial differential equations do not in general exist. Lumped electrical models and digital computer numerical models have been used to obtain approximate solutions. The research which in part is being reported was concerned with methods for systematically and efficiently finding small and accurate numerical models suitable for use in algorithms for computer-aided analysis and design of bipolar junction devices.

INTRODUCTION

The basic analytical model for describing the behavior of bipolar junction devices is characterized by the following basic semiconductor partial differential equations (PDE's) [1]

Poisson's Equation

$$\nabla^2 v = -\frac{q}{\epsilon}(c - n + p) \tag{1}$$

Hole Continuity Equation

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \vec{J}_p - R \tag{2}$$

Hole Current Density

$$\vec{J}_p = -q\mu_p p \nabla v - qD_p \nabla p \tag{3}$$

Electron Continuity Equation

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \vec{J}_n - R \tag{4}$$

---

[*] The author wishes to thankfully acknowledge support in part provided by AF-AFOSR Grant 1217-67 from the Applied Mathematics Group of the Air Force Office of Scientific Research, under Dr. D. A. Calahan. The work being reported was part of this author's Ph. D. thesis research, submitted to The University of Michigan.

Electron Current Density

$$\vec{J}_n = - q\, \mu_n\, n\nabla v + q\, D_n\, \nabla_n \qquad (5)$$

Total Current Density

$$\vec{J} = \vec{J}_p + \vec{J}_n - \epsilon \frac{\partial}{\partial t} (\nabla v) \qquad (6)$$

where R is the recombination rate,

c is the impurity doping distribution,

n is the electron carrier concentration,

p is the hole carrier concentration,

μ_p, μ_n are the mobilities, and

D_p, D_n are the diffusion constants,

with q and ϵ defined as in most general solid-state electronics texts. The analytical model conventionally adopted for semiconductor devices is the basic one dimensional equations. Even with this simplification and other approximations, analytic solutions are impossible to obtain under general conditions. To overcome this difficulty approximation techniques have been used which lead to approximate solutions to the analytic equations.

Early attempts to solve the equations were made by removing space variables from the semiconductor equations and modeling the semiconductor device with a lumped network model [2-4]. These models can be solved by digital circuit analysis programs; Linvill's diffusion model can use linear circuit analysis programs [2]; while Linvill's generalized model [2], Kani's model [3], and Ghosh's model [4] require nonlinear circuit analysis programs. Kani's model would appear to yield the closest approximation to the semiconductor device properties since his model is the only model developed from the partial differential equations without making any assumption as to the region of the device. Also, his model is closely related to the amount of charge stored in each lump with each lump representing a distributed region of the semiconductor device.

Numerical techniques, with the aid of digital computers, provide an alternative approach to the problem. Several general attempts have been made to solve the problem numerically [5-7]. Finite difference or differential-difference equations are used to approximate the semiconductor PDE's. The methods are generally applicable for forward bias and moderate injection levels, but are not readily applicable to high injection levels and two-dimensional problems. This author has found that most of these iterative techniques have a slow convergence rate. The accuracy of all these numerical solutions is dependent on the use of a large numerical model.

The major research effort has been concerned with the development of techniques for systematically finding the smallest numerical model which determines to a desired accuracy the electrical properties of a bipolar junction device. Consistent with this research effort, the mathematical tools developed provide for the systematic and efficient improvement of the accuracy of the numerical solution for a numerical model of any given fixed size.

Subordinate to this stated research effort is the numerical solution of the describing PDE's, referred to as the analysis phase. A different method for analysis has been developed for numerically solving the semiconductor equations. As in the previously reported methods for analysis, this method has accuracy problems and would seem to require large numerical models.

THE PROBLEM FORMULATION

Approaches using numerical techniques introduce discretization errors which adversely affects the accuracy of the approximate solution of Eqs. (1-6), and thus large numerical models are required to achieve good accuracy. In this investigation an optimization problem is defined which has as its objective function the reduction of the computational effort required to obtain a numerical solution, constrained by a specific accuracy criterion. The performance function can be defined quite generally in the modeling process. Possible definitions for the performance function require (1) the minimization of the truncation error or (2) a close match between the 'exact' and numerically computed terminal responses. The optimization parameters are the discrete lump sizes h_j, for j=1, 2, . . . , N; thus nonuniform lump sizes are being used to improve the accuracy of the numerical model. Other possible approaches for improving the approximate numerical solution are: (1) using a large number of uniform grid points while retaining the method with the same order truncation error, and (2) using a method with higher order truncation error while keeping N constant and uniform lumps. However these two approaches increase both the storage requirements and computational effort.

Thus a minimization problem has been defined. Mathematically, the minimization problem is the following:

$$\begin{aligned} &\text{Find minimum } N && (7)\\ &\text{such that } P = (P^* - \mathcal{P})^2 \leqq \alpha \\ &\qquad \underline{f}^k(\underline{u}^k, \underline{h}) = \underline{0} \\ &\qquad \underline{e}^T \underline{h} = h_1 + h_2 + \cdots + h_N = L_x \\ &\qquad h_j \geqq 0 \qquad j=1, 2, \ldots, N \end{aligned}$$

where P^* is the calculated performance,
$\mathcal{P}$ is the desired performance ,
N is the size of the numerical model,
P is the performance function ,
$\underline{f}^k = \underline{0}$ is the set of algebraic equations approximating the semiconductor equations,
$\underline{u}^k$ is the solution vector,
h_j are the lump sizes, j=1, 2, . . . , N,
α is the specified accuracy requirement , and
L_x is the length of the device.

This minimization problem has an integer objective function and a nonlinear constraint set, requiring a mathematical programming effort. The approach for solving this problem is presented briefly in outline form in the following section.

The Multi-Stage Process

For the minimization problem, Eq. (7) , a multi-stage process is proposed for finding the optimum N.

(1) Choose an initial value for N

(2) Use uniform lumps for the discretization

(3) Compute the discrete solution

(4) Minimize $P(\underline{h})$ (8)

subject to

$$\underline{f}^k(\underline{u}^k, \underline{h}) = \underline{0}$$

$$\underline{e}^T \underline{h} = L_x$$

$$h_j \geqq 0 \qquad j=1, 2, \ldots, N.$$

(5) Test if accuracy criterion is satisfied

(a) If $P \leqq \alpha$, decrease N

(b) If $P > \alpha$, increase N

(6) Repeat steps two through five until N is such that minimizing P for N lumps, $P \leq \alpha$; whereas minimizing P for N-1 lumps, $P > \alpha$.

The Minimization Procedure

The important step of this multi-stage process is step four which defines a nonlinear programming problem. Step five is relatively unimportant since it can be solved easily by any one of several systematic and efficient search procedures.

The nonlinear programming problem in the optimization process can be solved efficiently using a gradient minimization procedure. A very efficient gradient computation procedure has been developed which results in making the approach of nonuniform grid points feasible. This gradient procedure is in the class of adjoint methods [8, 9] and is applicable to general numerical formulations - including transient and higher dimensional problems.

The gradient computation procedure will be reported later. Some features of the procedure are:

(1) the method is directly applicable to both transient and d. c. problems, making it a unified gradient computation procedure,

(2) the method is applicable to both discrete and continuous formulations for digital computations,

(3) the accuracy of the gradient computation is directly dependent on the accuracy of the forward transient integration of the problem equations,

(4) the computational effort required to obtain the adjoint vector is less than or equal that for one forward integration of the problem equations,

(5) the initial condition for the adjoint vector at the final time is defined exactly by the computation procedure,

(6) this adjoint formulation is eighty times more efficient than the formulation using the chain rule from calculus.

The minimization procedure used was the Fletcher-Powell procedure [10]. The gradient vector was computed using the above outlined procedure and then modified by subtracting the last element of the vector from the other elements. The variables were kept positive by minimizing in the logarithm of the lump sizes which also has the effect of improving the convergence rate. With these modifications, the first (N-1) lump sizes are optimized using the Fletcher-Powell procedure, and the last llmp size is obtained from

$$h_N = L_x - \sum_{i=1}^{N-1} h_i \quad . \tag{9}$$

AN EXAMPLE

An example to illustrate the matching of the computed and 'exact' terminal responses will now be presented. The application considered is that of a turn-on transient in a short-base bipolar type diode. The terminal voltage is the input and the terminal current is the computed response. The example is implicitly an illustration of the use of non-uniform grid points as a technique for improving the accuracy of the numerical model.

The voltage transient being modeled discretely is as shown in Fig. 1 with the numbered points denoting the points used as discrete inputs. The 'exact' current densities were computed using a numerical model of 100 uniform lumps and will be denoted as $\tilde{J}_T^{\,k}$, k=0, 1, 2, 3, 4, corresponding to the numbers on Fig. 1. The computed current densities were obtained using a model with 20 lumps and will be denoted as $J_T^{\,k}$, k=0, 1, 2, 3, 4. The performance function was defined such that the error displayed is the percentage error difference in the computed and exact current densities. Since for the thermal equilibrium case the computed current density is identically equal the exact current density of $J_T^{\,k} = 0$, the first point, k = 0, is eliminated from the performance function.

The minimization problem from step four of the multi-stage process is:

Minimize

$$P(\underline{h}) = 10^4 \sum_{k=1}^{4} \left(\frac{1}{\tilde{J}_T^{\,k}} \right)^2 (J_T^{\,k} - \tilde{J}_T^{\,k})^2 \qquad (10)$$

subject to

$$\underline{f}^{\,k} = \underline{0}, \qquad k=0, 1, 2, 3, 4,$$

$$h_j \geqq 0, \qquad j=1, 2, \ldots, 20, \text{ and}$$

$$\sum_{j=1}^{20} h_j = 1.0$$

for the lump sizes scaled to make $L_x = 1.0$. Using the minimization procedure briefly discussed earlier, the results shown in Table 1 were obtained for $P(\underline{h})$ with the initial uniform lump sizes, and final nonuniform lump sizes and associated gradient vector elements shown in Table 2. These results were obtained for a device with a linearly graded impurity distribution of gradient 1.0×10^{20}. Thus a symmetric device was used and the results are for only one side of the device with lump number one at the metallurgical junction. The CPU time used may seem very costly; however, the time of 161.557 seconds for minimization for the twenty lump model is only twice as much time used for one analysis of the 100 lump model. As expected, the last lump size is large and is completely out of the space charge region which extends to x ≅ 0.548 maximum.

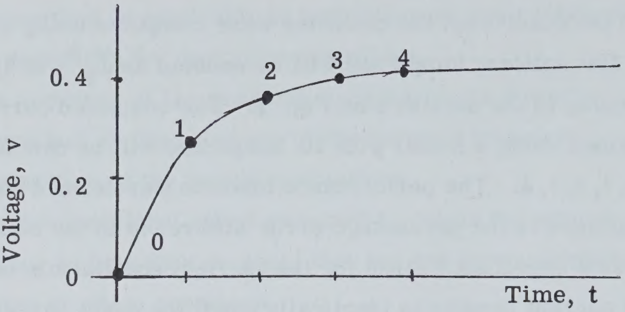


Figure 1 The Voltage Transient

Table 1

No. Gradient Computations	No. Evals. of P(h)	CPU Time (seconds)	P(h)
1	1	20.425	3252.900
2	7	94.751	56.223
3	15	142.927	11.045
4	20	161.557	10.995

Table 2

Lump No.	Initial Values		Final Values	
	Size	Gradient	Size	Gradient
1	0.05	16227	0.031489	14.531
2	0.05	11956	0.023608	- 4.7737
3	0.05	11535	0.028037	-15.794
4	0.05	11425	0.030745	-13.848
5	0.05	11377	0.030748	- 2.9906
6	0.05	11386	0.030624	4.6075
7	0.05	11378	0.030599	8.7584
8	0.05	11318	0.030628	11.098
9	0.05	11260	0.030667	12.538
10	0.05	11201	0.030711	13.516
11	0.05	11164	0.030874	14.230
12	0.05	11136	0.030868	14.786
13	0.05	11109	0.030848	15.228
14	0.05	11108	0.030822	15.590
15	0.05	11057	0.030773	15.903
16	0.05	10731	0.030650	16.215
17	0.05	10397	0.030270	16.662
18	0.05	10714	0.029184	17.286
19	0.05	10380	0.024359	17.941
20	0.05	2680	0.43350	10.420

CONCLUSIONS

The example presented illustrates the feasibility of using optimized nonuniform lump sizes to improve the accuracy of numerical models of semiconductor devices. This type of modeling effort provides the user with techniques for large signal modeling efforts. Other optimization examples minimizing the truncation error further indicate a great improvement in the overall accuracy of the model. The results obtained indicate and confirm that the density of the grid points in the numerical model should be greatest in the space-charge region (the space-charge region is the electrically active region for bipolar devices with the neutral region being essentially a bulk resistive element). These optimization techniques have also been shown to be directly applicable to optimizing two-dimensional numerical models.

REFERENCES

[1] Van Roosbroeck, W., "Theory of the Flow of Electrons and Holes in Germanium and Other Semiconductors", Bell Sys. Tech. Journal, Vol. 29, No. 4, pp. 560-607, Oct. 1950.

[2] Linvill, J. G., Models of Transistors and Diodes, McGraw-Hill, New York, 1963.

[3] Kani, K. and Ohtsuki, T., "A Unified Modeling Scheme for Semiconductor Devices with Applications of State Variable Analysis", IEEE International Symp. on Circuit Theory, Dec. 1968 and IEEE Trans. Circuit Theory, 1969.

[4] Ghosh, H. N., "A Distributed Model of the Junction Transistor and Its Application in the Prediction of the Emitter-Base Diode Characteristic, Base Impedance, and Pulse Response of the Device", IEEE Trans. Electron Devices, Vol. ED-12, No. 10, pp. 513-531, Oct. 1965.

[5] Gummel, H. K., "A Self-Consistent Iterative Scheme for One-dimensional Steady State Transistor Calculations", IEEE Trans. Electron Devices, Vol. ED-11, pp. 455-465, Oct. 1964.

[6] DeMari, A., "An Accurate Numerical One-dimensional Solution of the p-n Junction Under Arbitrary Transient Conditions", Solid-State Electronics, Vol. 11, No. 11, pp. 1021-1053, Nov. 1968.

[7] Scharfetter, D. L., and Gummel, H. K., "Large-Signal Analysis of a Silicon Read Diode Oscillator", IEEE Trans. Electron Devices, Vol. ED-16, No. 1, pp. 64-77, January 1969.

[8] Hachtel, G. D., and Rohrer, R. A., "Techniques for the Optimal Design and Synthesis of Switching Circuits", Proc. IEEE, Vol. 55, No. 11, pp. 1864-1877, Nov. 1967.

[9] Director, S. W. and Rohrer, R. A., "Automated Network Design - The Frequency Domain Case", IEEE Trans. Circuit Theory, Vol. CT-16, pp. 330-337, August 1969.

[10] Fletcher, R., and Powell, M. J. D., "A Rapidly Convergent Descent Method for Minimization", Computer J., Vol. 6, pp. 163-168, July 1963

MINIMIZATION OF SENSITIVITY OF ELECTRICAL NETWORKS OVER A FREQUENCY INTERVAL

Virendra K. Manaktala
Bell Telephone Laboratories
Indianapolis, Indiana 46206

Gerry L. Kelly
Department of Electrical Engineering
University of Kansas
Lawrence, Kansas 66044

ABSTRACT

Computational methods are presented for direct synthesis of electrical networks subject to integral sensitivity measures and practical constraints on the network parameters. Sensitivity measures which are valid over the entire parameter constraint space and frequency interval are considered instead of those restricted to a small neighborhood of a single point in parameter constraint space or a single frequency.

I. INTRODUCTION

Most of the well known definitions of single and multiparameter sensitivity consider only incremental changes in the network elements. However, the present technology in integrated circuit fabrication requires the consideration of large changes in the element values. Moreover, the existing sensitivity definitions are given for a single frequency only and are meaningful for network comparison only if one of the networks has lower sensitivities than the other networks at all frequencies. In this paper a multiparameter sensitivity definition which considers large changes in the element values over a desired band of frequencies is utilized to compare the worst case sensitivities of two given networks and improving the worst case sensitivity of a chosen realization.

II. A NEW DEFINITION OF MULTIPARAMETER SENSITIVITY

The definition of classical sensitivity [1] and definitions of multiparameter sensitivity given by Goldstein and Kuo [2], Blostein [3], Remec [4], etc. are well known. These consider small deviations in element values and compute the sensitivity at a single frequency. Hakimi and Cruz [5] defined multiparameter sensitivity at a frequency ω_k as the maximum of the actual normalized percentage change in a network function divided by the product of the element tolerances where the element values are constrained within these tolerances.

$$S^{T(j\omega_k)}_{x_1,x_2,\ldots,x_n} = \frac{\mathrm{Max}\,\Delta T(j\omega_k)}{T(j\omega_k)[\delta_1\cdot\delta_2\cdot\ldots\cdot\delta_n]} \tag{1}$$

$$0 \le |\epsilon_i| \le \delta_i$$

where δ_i is the tolerance on the element x_i and ϵ_i is the actual per unit deviation $\Delta x_i/x_i$. Definition (1) seems quite appropriate since it considers large changes in the element values; however, it yields a different value of sensitivity at each frequency for any particular network which means that it will be a useful network selection criterion only when one network has lower sensitivity at all frequencies in the desired frequency band than another network. Since this is not always the case, it is desirable to have a sensitivity measure which gives a meaningful comparison of different realizations over the entire frequency interval being considered. Kelly [6] has extended Hakimi and Cruz's definition to yield a measure of network multiparameter sensitivity over a frequency band rather than at a single frequency. This definition for multiparameter sensitivity for a frequency interval $[\omega_a, \omega_b]$ is

$$S_X^T = \int_{\omega_a}^{\omega_b} \lambda(\omega)[T_{max} - T_{min}]d\omega \tag{2}$$

where $T_{max}(\omega)$ and $T_{min}(\omega)$ are respectively the maximum and minimum values that the network function $T(\omega)$ can obtain for element values

$$R = [x_i \mid (1-\delta_i)x_{iN} \leq x_i \leq (1+\delta_i)x_{iN}] \tag{3}$$

contained in the element constraint set, R.The factor $\lambda(\omega)$ can be the desired weighting for a particular frequency range for a specific application or a factor to compensate for a scale factor difference in the event that two networks realize the same poles and zeros with a different constant multiplier or a combination of these factors. When $\lambda = 1$, the definition reduces to the area between the actual T_{max} and T_{min} curves. This definition is not normalized by the network function value at each frequency so as to avoid overweighting certain parts of the frequency band. Hence knowing T_{max} and T_{min} at different frequencies, multiparameter sensitivity over a band of frequency can be obtained.

III. FORMULATION OF THE PROCEDURE

An optimization procedure done on a digital computer is presented which gives us a convenient tool for comparing the sensitivity of two or more realizations of the same transfer function, over a band of frequencies of interest. It can also be used as a sensitivity analysis technique after a realization has been decided upon. In this case it will provide a valuable insight into the frequency response of the network when all parameters are simultaneously subjected to large tolerances. The method consists of finding the maximum and minimum values of the network function over the entire element constraint set at discrete frequencies in the band of interest when all the parameters are allowed to undergo large changes and vary simultaneously. The tolerances on the elements are known and specified, in addition, to some realistic constraints on the values of network elements. Definition (2) is then utilized to calculate the sensitivity over the same band of frequencies and same tolerances on the elements of different realizations. The network chosen would be the one with smallest sensitivity of all network configurations being compared.

The problem is converted into a nonlinear programming problem which can be stated as follows: Determine the vector X* which minimizes (maximizes) a network function $T(\omega, X)$ at a set of discrete frequencies, ω_j, $j = 1,2,\ldots,m$, subject to constraints

1. each $x_k > 0$ (4)
2. $x_{kN}(1-\delta_k) \leq x_k \leq x_{kN}(1+\delta_k)$, $k = 1,2,\ldots,n$ (5)

where δ_k is the tolerance on element x_k and x_{kN} is its nominal value. In this case we have a nonlinear objective function with linear constraints. A transformation suggested by Carroll [7] and proved by Fiacco and McCormick [8] is utilized to convert the constrained optimization problem into a sequence of unconstrained optimization problems. T_{max} and T_{min} are calculated over the entire element constraint set rather than at the extreme element values. Also T_{max} and T_{min} are the actual function maxima and minima respectively rather than just being conservative bounds as given by Hakimi and Cruz. Also their method required 2^n computations of the network function for an n element network which runs into an unpractical amount of computation time. Their bounds are quite conservative in view of the fact that they were not obtained by element values contained in element constraint set R. Fiacco and McCormick [9] and Spang [10] have shown that the speed of convergence is highly dependent upon the accuracy to which the partial derivatives can be determined. These partials were generated using topological techniques in order to avoid excessive error when the network function is approximated by the first three terms of the Taylor series. The network function considered for the comparing two realizations is the magnitude squared voltage ratio function, although any other network function could be used. These were determined topologically by extending the algorithm originally due to Mayeda [11,12]. The information about the network coefficients and the first and second partial derivatives are obtained utilizing a one time analysis program OTAN [13]. As long as the network can be analyzed topologically the partials can be calculated accurately regardless of the size of the networks. This is quite attractive because optimization can be done with respect to a large number of parameters. The details of the formulation of the network function and its derivatives are given.

IV. EXAMPLE OF COMPUTATIONAL PROCEDURE

It is well known that for a doubly terminated coupling network, the zeros of reflection coefficients may be selected in either the right half of the complex frequency S-plane (RHP) or the left half plane (LHP) or a combination of zeros in the right half plane and laft half plane. It can be easily shown that for polynomial filters $2^{(n-1)/2}$ and $2^{n/2-1}$ different combinations of pairs of zero distributions exist for n odd and even, respectively (where n represents the degree of the denominator polynomial of the transfer function). Weinberg [14] has obtained the element values for Chebyshev filters for zeros of the reflection coefficient either all in the left half plane or alternating between the left and right half planes. Network 1 is a 5th-order normalized Chebyshev filter with 1/4 db pass band ripple having the zeros of the reflection coefficient alternating between the left and right half planes [14]. Network 2 is also a 5th-order normalized Chebyshev filter with 1/4 db pass band ripple; however, the zeros of the reflection coefficient are all in the left hand plane[14]. The computational method was utilized to find the actual $T_{max}(\omega_k)$ and $T_{min}(\omega_k)$ where $T(\omega_k)$ is the magnitude of the voltage ratio function. The values of T_{max} and T_{min} are obtained to find and compare the sensitivity of two realizations when the elements are subjected to a fixed tolerance. In each case the weighting factor λ is set equal to 1 and therefore, the area between T_{max} and T_{min} curves gives us the sensitivity of different realizations over the normalized band of frequencies. The graphs of worst case performance when the element values are subjected to a tolerance of 20 per cent are shown in Figures 1 and 2. It is seen that the network realizations whose reflection coefficient zeros alternate in the LHP and RHP are more sensitive to parameters changes than the networks having reflection coefficient zeros in the LHP only, when the specified worst case tolerance on the network parameter is the same in both cases. Other examples have also been run and the results are tabulated in [27].

V. SYNTHESIS OF MINIMUM SENSITIVITY NETWORKS

A method of analyzing the worst case sensitivities over a band of frequencies has been presented. Often the comparison of sensitivities on a worst case basis or any other basis is used as a network selection criterion. It is clear that although many definitions of network sensitivity have been proposed and used for sensitivity analysis, few synthesis procedures which approximate a given network function with minimum sensitivity have been reported in literature. Schoeffler [16,17] used the theory of equivalent networks to generate a sequence of networks whose transfer functions are identical to that of a given network but whose elements differ from one network to the next by an incremental amount. As a measure of the optimality of the network, the sum of squared magnitudes of the sensitivies summed over the network elements was used as a performance criterion. After the networks were generated, the performance criterion was applied to them and the network with the lowest value of sensitivity is the one chosen. This synthesis procedure is not optimum since all possible continuous equivalent networks are not generated [18] and the values of network elements which are generated are not controllable. This can result in unrealistic values of the network elements. The sensitivity criterion is good only for one frequency although it has been observed by Leeds and Ugron [19] that the network with the minimum total sensitivity at a frequency is also nearly optimum at another frequency. Blostein [20] has modified the performance criterion so that it can be used over a certain range of frequencies. Schmidt and Kasper [21] and Kuh and Lau [22] have derived some properties of the sensitivity of equivalent networks.

Although continuous equivalent network theory apparently has a lot of potential in the synthesis of networks, it has not so far been utilized. A recent paper by Shirakawa et al. [23] is, in the author's opinion, an excellent attempt at utilizing continuous equivalent network theory for synthesis of minimum sensitivity networks. The paper describes a synthesis procedure for passive networks with minimum sensitivity to element tolerances through some classes of continuously equivalent transformations. The method of transformations discussed are those transformerless ones in which the port structures of networks are invariant. Two approaches to the problem are considered. The first is the procedure in which an additional network is sought such that the augmented network, obtained by parallel connection of the given network and the additional one, is equivalent to the original one. The second is based on some well known transformations [24]. Then, given a continuous equivalent network, the next step is the choice of transformation parameters which minimze sensitivity to element tolerances. This is accomplished through the penalty function approach described in [8]. The constraints are the non-negativity of elements and the function minimized is the maximum sensitivity at a single frequency, over the entire band of frequencies. The results are as expected. The total sensitivity decreases with the increase in the number of elements and the network with minimum total sensitivity at a frequency has nearly minimum sensitivity at other frequencies also.

In the next section a method is presented which is more general and versatile than any of the methods described above. It does not use the continuous equivalent theory and is applicable to both active and passive networks.

VI. A METHOD TO SYNTHESIZE NETWORKS WITH MINIMUM SENSITIVITY

In the previous section, synthesis of minimum sensitivity networks using continuous equivalent theory was described. A more general method utilizing nonlinear programming is presented which offers considerable advantages over the existing methods. The problem of realizing networks with minimum sensitivity to element tolerances is very important and a synthesis procedure applicable to actual design of networks is required. In many cases the network structure is already known or defined due to the nature of job the network is supposed to perform. The upper and lower bounds to element values are usually limited to certain values, for example, in integrated circuits the resistors and capacitors can only have a certain range of values. Unrealistic element values are useless and the constraints on them should be incorporated in any synthesis procedure. The method should also have the flexibility of incorporating element ratio constraints which are important in the modern circuit fabrication technology. Keeping these things in mind, a definition for the problem of synthesizing practical networks with minimum sensitivity is now given.

Definition 1

Given a network (active or passive) synthesized by any of the known schemes, find the element values of the desired network such that the new network has the same topology and transfer function as the original one, realistic constraints on element values and for element ratios are met and the sensitivity of the network function to the element deviations (due to their tolerances or other reasons) is minimized in accordance with a known error criterion.

The synthesis problem defined above is solved as a modified nonlinear programming problem with both inequality and equality constraints. The classic NLP problem with inequality and equality constraints can be stated as: Determination of a vector X* which minimizes f(X) subject to

$$g_i(X) \geq 0 \qquad i = 1, 2, \ldots, m$$

and

$$h_j(X) = 0 \qquad j = 1, 2, \ldots, p$$

where f(X) is the function to be minimized, $g_i(X)$ are the inequality constraints as defined in equations (4-5), $h_j(X)$ are the equality constraints. This approach is clearly feasible to solve the synthesis problem outlined in Definition 1. The particular transformation suggested and proved by Fiacco and McCormick [25] for linear equality constraints is

$$P(X_k, r_k) = f(X_k) + r_k \sum_{i=1}^{m} 1/g_i(X_k) + \sum_{j=1}^{p} h_j^{\,2}(X_k)/r_k^{\,1/2} \tag{6}$$

where all the quantities in equation (6) have been previously defined except $h_j(X_k)$.

$h_j(X) = 0$ is equivalent to

$$-h_j(X) \geq 0 \tag{7}$$

$$h_j(X) \geq 0 \tag{8}$$

In the initial approach tried for solving minimum sensitivity synthesis problem, f(X) was the actual worst deviation from nominal value of the transfer function, at a single frequency over the entire frequency range Ω.

$$f(X) = \max_{\omega \epsilon \Omega} \left\{ \left| T_{max} - T_{nom} \right| \text{ or } \left| T_{min} - T_{nom} \right| \right\} \tag{9}$$

where all the quantities in equation (9) have been defined previously, $g_i(X)$ are the inequality constraints and $h_j(X)$ are the constraints formed from the coefficients of the nominal transfer function (both numerator and denominator) and the current value of these coefficients during the optimization procedure.

$h_j = a_j - c_j$ where

a_j = actual coefficient

c_j = nominal coefficient

$$\sum_{j=1}^{p} h_j = \sum_{j=1}^{\ell+m} (a_j - c_j) \tag{10}$$

ℓ = number of numerator coefficients
m = number of denominator coefficients

It was found after trying many possibilities for the initial value of r_k and other numerous factors, the transformation in equation (6) did not converge to a solution such that

$$f = P = G \tag{11}$$

where G is an upperbound to the solution or a solution to the dual problem

$$G = f(X_k) - r_k \sum_{i=1}^{m} 1/g_i(X_k) + 2\sum_{j=1}^{p} h_j^2(X_k)/r_k^{1/2} \tag{12}$$

It is clear that equation (11) is satisfied only when

$$\sum_{j=1}^{p} h_j^2(X)/r_k^{1/2} = 0 \tag{13}$$

and

$$r_k \sum_{i=1}^{m} 1/g_i(X) \leq \epsilon \tag{14}$$

where ϵ is a very small number of the order of 10^{-10}. The reason for this nonconvergence could be that this technique is formulated for linear equality constraints only and the convergence criteria for nonlinear equality constraints are not known. It is well known that the coefficients of the network function are multilinear functions of the network elements. Also no suitable method for decreasing r_k systematically is known, such that convergence will be guaranteed. When $r_k < 1$, the equality constraints are overweighted and the vector of elements which decreases the inequality constraints and the function as $r_k \rightarrow 0$, actually increases the equality constraints and the solution diverges after a certain point. It became apparent that two different "r_k's" have to be used. Different values were tried and the best weighting for equality constraints was found to be 1. The modified transformation used is

$$P(X_k) = f(X_k) + r_k \sum_{i=1}^{m} 1/g_i(X_k) + \sum_{j=1}^{p} h_j^2(X_k) \tag{15}$$

The only change from equation (6) is that equality constraints are not weighted by $1/r_k^{1/2}$. As the inequality constraints decrease with r_k, the function $f(X_k)$ and $h_j^2(X_k)$ also decrease. If convergence is not reached after $r_k \rightarrow 0$, an iterative procedure is used which perturbs the element values until the actual coefficients are within the desired closeness to the nominal transfer function coefficients. This decreases the equalities constraints further such that convergence is reached. Also, we can get as close to the nominal transfer function as desired. The details of this iterative perturbation procedure are given in [27]. It is found that this modification of the classical NLP problem for minimum sensitivity synthesis of electrical network works quite well. A numerical example which illustrates the synthesis procedure described above is described in detail in the next section.

VII. EXAMPLE OF THE SYNTHESIS METHOD

The numerical example chosen for illustrating the procedure described in the last section is shown in Figure 3. It is a three stage negative feedback amplifier whose element values have been normalized. Since this problem has 12 elements, three of which are active, it is felt that it is a good example to illustrate the minimum sensitivity synthesis procedure. Given this realization, it is desired to achieve, within a multiplicative constant, an identical transfer function to the nominal transfer function, so that the worst case sensitivity over the entire frequency interval when all parameters are allowed to undergo large changes is reduced.

The first step was to run a worst case sensitivity analysis on it. This was done while all the elements were permitted to undergo ± 20 per cent change in their element values. The worst deviation from the nominal occurred at a frequency of .1778268 where $T_{max} = 6.50532413$ and $T_{nom} = .2936732$, i.e., $T_{max} - T_{nom} = 6.21165093$. This is the sensitivity function which will be minimized. T_{nom} is read in as a constant and hence only the partial derivatives of T_{max} are required to be calculated. T, the network function in this case is the voltage ratio function squared.

$$f(X) = \left| T_{max} - T_{nom} \right| \tag{16}$$

$$\therefore \frac{\partial f(X)}{\partial X_i} = \frac{\partial T_{max}}{\partial X_i} \tag{17}$$

The conjugate gradient method is used and hence only the first order partial derivatives are required. The constraints on this optimization problem are both of equality and inequality type. Equality constraints are formulated in terms of the network function coefficients (both numerator and denominator), so that the transfer function of the new network and the original network are the same. The numerator and denominator coefficients of the network function are read in and the current value of these coefficients are evaluated during the optimization procedure. Typical equality constraints for a single coefficient a_i are formulated as

$$h_{i1} = a_i - c_i \tag{18}$$

$$h_{i2} = c_i - a_i \tag{19}$$

It is clear that equation (18) and (19) will be only satisfied when $a_i = c_i$ where h_{i1} and h_{i2} = equality constraints for a_i

a_i = actual network function coefficient

c_i = nominal network function coefficient

The number of equality constraints is twice the number of network coefficients.

The inequality constraints ensure the non-negativity of the network elements and keep the element values within a specified range. The number of inequality constraints for this example is 24, corresponding to a high and low value for each element. For the problem being considered, the voltage ratio function had three numerator coefficients and four denominator coefficients and hence there were fourteen equality constraints. The values of elements which form T_{max} are used as the initial starting vector.

The computer program "MINMAX" used for this problem is a modified version of program "MINIMUM ERROR" described in detail in [15]. A listing of MINMAX may be found in [27]. The function is to be minimized, that is, $T_{max} - T_{nom}$ is reduced from an original value of 6.211 to .07. A worst case analysis was performed on both the original and the new network with ± 20 per cent tolerance on the elements. A comparison of sensitivities shows that the sensitivity of the new realization is lower by 21 per cent. The graphs of the worst case deviation for the original and the new realization are plotted in Figure 4 and 5 respectively.

VIII. CONCLUSIONS

A method has been presented which gives us a meaningful criterion to choose a realization when a network function can be synthesized by more than one realization. Different networks which have the same transfer function can be the ones with different distribution of reflection coefficient zeros, with series parallel or nonseries parallel realization as in [26] or equivalent networks as in [18]. In the past sensitivity criteria have been defined for one frequency which are only meaningful when one realization has lower sensitivity than the other one at all frequencies which is very seldom the case. Hence the method presented which considers the sensitivity over an entire band of frequencies is a definite improvement over the previous sensitivity analysis techniques. The method can also be used for analyzing multiparameter sensitivity when only one realization is given and one is interested in the frequency response of a network function when all network parameters are subject to large changes in their nominal values simultaneously. A procedure for synthesizing network with minimum worst case sensitivity is given in the second part of the paper. The sensitivity criterion used is the worst case multiparameter sensitivity as described in the first part of this chapter. This approach of synthesizing network while optimizing the sensitivity is quite general compared to the existing techniques which invariably make use of continuous equivalent transformations [24]. It should be noted that the actual network elements and not the transformation parameters are found which give an optimum network with regard to sensitivity. Also only the optimum network with controlled element values is realized unlike [18] where many equivalent networks are generated and then tested for optimality with regard to a performance criterion at a single frequency. This saves computational time in addition to the saving effected through the use of an efficient gradient technique. The average execution time for a single optimization (maximization or minimization) for the numerical example discussed above was twenty-one seconds.

REFERENCES

1. Bode, H. W., Network Analysis and Feedback Amplifier Design, D. Van Nostrand Co., Inc., Princeton, New Jersey, 1945.
2. Kuo, F. F., "Pole-zero Sensitivity in Network Functions," IRE Trans. on Circuit Theory, vol. CT-5, no. 4, p. 372, December 1958.
3. Blostein, M. L., "Some Bounds on Sensitivity RLC Networks," Proc. First Allerton Conf. on Circuit and System Theory, pp. 488-501, November, 1963.
4. Remec, M. J., "Application of Multiparameter Sensitivity in Linear Systems," Doctoral dissertation, University of Illinois, 1963.
5. Hakimi, S. L. and J. B. Cruz, Jr., "Measures of Sensitivity for Linear Systems with Large Multiparameter Variation," IRE Wescon Convention Record, Part 2, pp. 109-15, August 1960.
6. Kelly, G. L., "A Measure of Multiparameter Sensitivity," Proc. 12th Midwest Circuit Theory Symposium, April 1969, Austin, Texas.

7. Carroll, C. W., "The Created Response Surface Technique for Optimizing Nonlinear Restrained Systems," Operations Res., vol. 9, no. 2, pp. 169

8. Fiacco, A. V. and G. P. McCormick, "Programming Under Nonlinear Constraints by Unconstrained Minimization: A Primal-Dual Method," Research Analysis Corporation, RAC-TP-96, September 1963.

9. Fiacco, A. V. and G. P. McCormick, "The Sequential Unconstrained Minimization Technique for Nonlinear Programming, Algorithm II, Optimum Gradients by Fibonacci Search," Research Analysis Corporation, RAC-TP-123, June 1864.

10. Spang, H. A., III, "A Review of Minimization Techniques for Nonlinear Functions," SIAM Rev., vol. 4, no. 4, pp. 343-365, 1962.

11. Mayeda, W., "Topological Formulas for Active Networks," Interim Technical Report No. 8, U.S. Army Contract DA-11-02200RD-1983, University of Illinois, Urbana, Illinois, January 1957.

12. Mayeda, W., "Digital Determination of Topological Quantities and Network Functions," Interim Technical Report No. 6, US Army Contract DA-11-022-ORD-1983, University of Illinois, Urbana, January 1957.

13. Kelly, G. L., "Algebraic Generation of Network Coefficients and Partial Derivatives," Proc. 6th Allerton Conf. on Circuit and System Theory, pp. 843-844, October 1968.

14. Weinberg, L., Network Analysis and Synthesis, McGraw-Hill, Inc., New York, 1962.

15. Manaktala, V. K. and G. L. Kelly, "Application of Nonlinear Programming to Worst Case Sensitivity Analysis and Least Squared Error Synthesis of Electrical Networks," Tech. Report 145-4, Center for Research in Engineering Science, University of Kansas, Lawrence, January 1970.

16. Schoeffler, J. D., "The Synthesis of Minimum Sensitivity Networks," IEEE Trans. on Circuit Theory, vol. CT-11, no. 2, pp. 271-76, June 1964.

17. Schoeffler, J. D., "Continuously Equivalent Networks and Their Application," IEEE Trans. Communication and Electronics, vol. 83, pp. 763-767, November 1964.

18. Newcomb, R. W., "The Noncompleteness of Continuously Equivalent Networks," IEEE Trans. Circuit Theory (Correspondence), vol. CT-13, pp. 207-208, June 1966.

19. Leeds, J. V. and G. I. Ugron, "Simplified Multiple Parameter Sensitivity Calculation and Continuously Equivalent Networks," IEEE Trans. Circuit Theory, vol. CT-14, pp. 188-191, June 1967.

20. Blostein, M. L., "Generation of Minimum Sensitivity Networks," IEEE Trans. Circuit Theory (Correspondence), vol. CT-14, pp. 87-88, 1967.

21. Schmidt, G. and G. Kasper, "On Minimum Sensitivity Networks," IEEE Trans. Circuit Theory (Correspondence), vol. CT-14, pp. 438-40, 1967.

22. Kuh, E. S. and C. G. Lau, "Sensitivity Invariants of Continuously Equivalent Networks," IEEE Trans. Circuit Theory, vol. CT-14, pp. 175-77, September 1968.

23. Shirakawa, I., T. Temma, and H. Ozaki, "Synthesis of Minimum-Sensitivity Networks through Some Classes of Equivalent Transformations," IEEE Trans. Circuit Theory, vol. CT-17, no. 1, pp. 2-8, February 1970.

24. Newcomb, R. W., Linear Multiport Synthesis, McGraw-Hill, New York, 1966, Chapter 10.

25. Fiacco, A. V. and G. P. McCormick, "The Sequential Unconstrained Minimization Technique for Programming with Equality Constraints," Research Analysis Corporation, RAC-TP-155, April 1965.

26. Lee, S. C., "On Multiparameter Sensitivity," Third Allerton Conf. on Circuit and Systems Theory, October 1965.

27. Manaktala, V. K., "Minimization of Sensitivity Over a Frequency Interval and Synthesis of Electrical Networks with Practical Parameter Constraints," Ph. D. dissertation, The University of Kansas, May 1970.

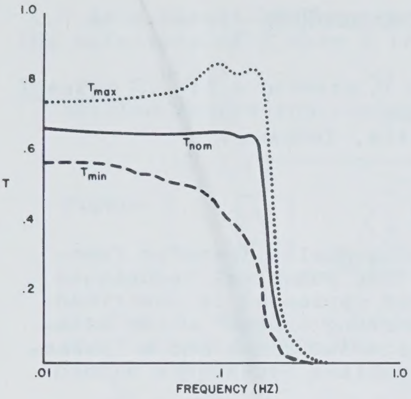


FIGURE 1 WORST CASE VARIATION FOR ± 20 PERCENT TOLERANCE ON NETWORK 1 PARAMETERS

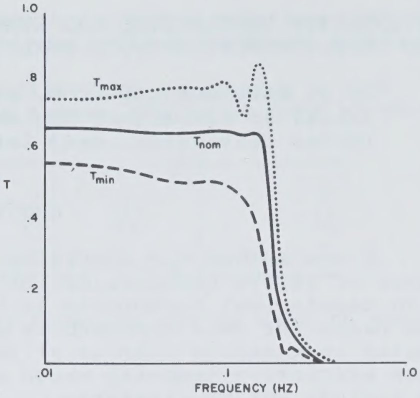


FIGURE 2 WORST CASE VARIATION FOR ± 20 PERCENT TOLERANCE ON NETWORK 2 PARAMETERS

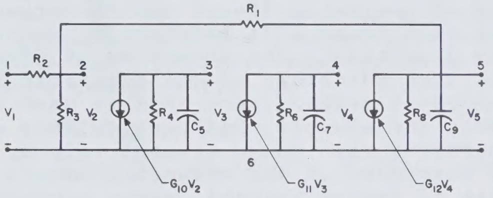

R_1
R_2
1
2
3
4
5
+
V_1
R_3
V_2
R_4
C_5
V_3
R_6
C_7
V_4
R_8
C_9
V_5
–
6
$G_{10}V_2$
$G_{11}V_3$
$G_{12}V_4$

FIGURE 3 A THREE-STAGE NEGATIVE FEEDBACK AMPLIFIER.

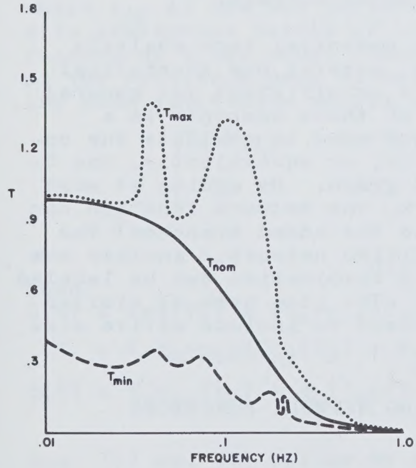


FIGURE 4 WORST CASE VARIATION OF THE ORIGINAL NETWORK WITH ±20 PERCENT TOLERANCE ON NETWORK ELEMENTS

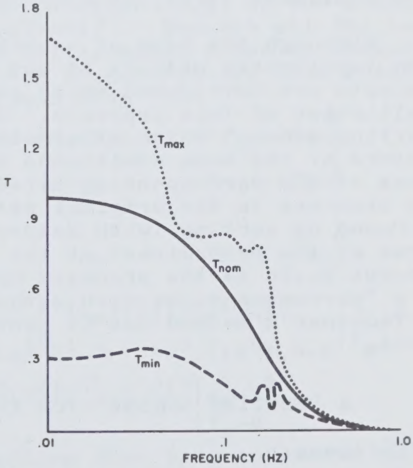


FIGURE 5 WORST CASE VARIATION OF THE OPTIMIZES NETWORK WITH ±20 PERCENT TOLERANCE ON NETWORK PARAMETERS

INTEGRATING TOPOLOGICAL AND NUMERICAL METHODS FOR SEMI-SYMBOLIC NETWORK ANALYSIS

G. E. ALDERSON AND P. M. LIN
School of Electrical Engineering
Purdue University, West Lafayette, Indiana

ABSTRACT

A new method for obtaining semi-symbolic transfer functions of active networks by integrating numerical techniques with topological techniques (a hybrid approach) is described. The basis for this approach is a "sorting scheme" which eliminates enumerating 2-trees of the directed graph and a "parameter extraction process" which generalizes Feussner's method to include active elements.

INTRODUCTION

A new method of analyzing linear active networks by integrating numerical techniques with topological techniques (a hybrid approach) is described. The objective of forming a hybrid combination is to take advantage of the desirable features of each method. Numerical methods such as that used in ECAP, NET-1, and SCEPTRE, for example, are particularly adept at handling large networks with known element values. However, if it is desired to keep some of the network elements in symbol form (for purposes of design insight, error control, parameter variation, optimization, etc.), the numerical techniques must be replaced by topological techniques. Unfortunately, the topological methods now available are basically inefficient in that they require evaluating all the trees of the graph, produce duplicate terms, or require finding k-trees.

Although the idea of combining numerical type analysis with topological methods is not new, several new theoretical concepts are introduced which permit an efficient yet general realization of this approach. One of these concepts is a "sorting scheme" which eliminates the need to evaluate the cofactors of the node admittance matrix, or equivalently, the 2-trees of the corresponding directed graph. By adding at most two branches to the original network, the network function can be found by sorting (with respect to the added branches) the terms of the determinant of the modified network. Another new concept basic to the proposed hybrid combination can be labeled as a "parameter extraction method." The idea here is similar to Feussner's method but is generalized to include active elements.

A "SORTING SCHEME" FOR FINDING NETWORK FUNCTIONS

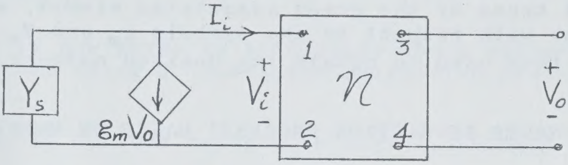
Definitions:
Y = nxn indefinite admittance matrix.
$C_{i\ell}(X)$ = cofactor of element (i,ℓ) of the matrix X.

Theorem 1: All the cofactors of Y are equal.
Proof: See reference [1].

As a result of Theorem 1, we can use the notation C(X) for the cofactors of X when X is an indefinite admittance matrix.

Theorem 2: If a network $\mathcal{N}$ is modified by connecting a controlled source and admittance across the input port as shown

Figure 1



in Fig. 1, then the following formulas can be used to obtain the network functions:

$$\frac{V_o}{V_i} = \frac{P_{g_m}}{P_{Y_s}}, \qquad \frac{V_o}{I_i} = \frac{P_{g_m}}{P_o}, \qquad \text{and} \quad \frac{V_i}{I_i} = \frac{P_{Y_s}}{P_o} \tag{1}$$

where $C(Y) = Y_s P_{Y_s} + g_m P_{g_m} + P_o$ (2)

and $C(Y)$ = determinant of the node admittance matrix representing the complete network consisting of $\mathcal{N}$ together with the two added branches.

Proof: Consider the indefinite admittance matrix, Y, of the complete network shown in Fig. 1.

$$Y = \begin{bmatrix} Y_s+y_{11} & -Y_s+y_{12} & g_m+y_{13} & -g_m+y_{14} & y_{15}\cdots y_{1n} \\ -Y_s+y_{21} & Y_s+y_{22} & -g_m+y_{23} & g_m+y_{24} & y_{25}\cdots y_{2n} \\ y_{31} & y_{32} & y_{33} & y_{34} & y_{35}\cdots y_{3n} \\ & & \vdots & \cdot & \\ y_{n1} & y_{n2} & y_{n3} & y_{n4} & y_{n5}\cdots y_{nn} \end{bmatrix}$$

where $y_{i\ell}$ is the set of terms at position (i,ℓ) of the indefinite admittance matrix of the network $\mathcal{N}$. Because all the cofactors of Y are equal (Theorem 1), terminal 2 can be chosen as the reference node without loss of generality. The resulting node admittance matrix as found from Y is

$$Y_2 = \begin{bmatrix} Y_s+y_{11} & g_m+y_{13} & -g_m+y_{14} & y_{15}\cdots y_{1n} \\ y_{31} & y_{33} & y_{34} & y_{35}\cdots y_{3n} \\ & & \vdots & \\ y_{n1} & y_{n3} & y_{n4} & y_{n5}\cdots y_{nn} \end{bmatrix}$$

Then

$$C(Y) = \det[Y_2] = (Y_s+y_{11})C_{11}(Y_2) + (g_m+y_{13})C_{12}(Y_2) + (-g_m+y_{14})C_{13}(Y_2) + y_{15}C_{14}(Y_2) + \ldots + y_{1n}C_{1,n-1}(Y_2)$$

$$C(Y) = Y_sC_{11}(Y_2) + g_m\left[C_{12}(Y_2)-C_{13}(Y_2)\right] + C(Y\Big|_{\substack{g_m=0\\Y_s=0}})$$

Eqs. (1) and (2) follow by recalling that (see reference [2])

$$\frac{V_o}{V_i} = \frac{C_{12}(Y_2)-C_{13}(Y_2)}{C_{11}(Y_2)}, \quad \frac{V_o}{I_i} = \frac{C_{12}(Y_2)-C_{13}(Y_2)}{C(Y\Big|_{\substack{g_m=0\\Y_s=0}})}, \quad \frac{V_i}{I_i} = \frac{C_{11}(Y_2)}{C(Y\Big|_{\substack{g_m=0\\Y_s=0}})}$$

and defining

$P_{Y_s} = C_{11}(Y_2)$, $P_{g_m} = C_{12}(Y_2)-C_{13}(Y_2)$, and $P_o = C(Y\Big|_{\substack{g_m=0\\ Y_s=0}})$ QED

To apply Theorem 1 we find C(Y), or equivalently, enumerate all directed trees of the graph associated with Y, and then sort the terms with respect to the symbols g_m and Y_s. Eqs. (1) and (2) are then used to obtain the desired network functions.

A "PARAMETER EXTRACTION PROCESS" BASED ON MATRIX MODIFICATION

Consider a network containing R, L, and C type elements and voltage controlled current sources (referred to as an active element in this paper). A general element of the network can be represented in the indefinite admittance matrix by the terms

$$\begin{array}{lll} \alpha \text{ in position } (i,\ell) & & \\ -\alpha \text{ in position } (i,m) & i < j & \\ -\alpha \text{ in position } (j,\ell) & \ell < m & \\ \alpha \text{ in position } (j,m) & & \end{array} \qquad (3)$$

Refer to the indefinite admittance matrix Y, next page, for an example of a passive type element Y_s $(i=\ell=1, j=m=5)$ and an active type element $g_a(i=3, \ell=2, j=m=4)$.

The following theorem gives the theoretical basis by which the parameter α associated with a general element (i.e., active or passive) can be extracted from the cofactors of Y. Define X_{km} as the matrix formed from X by deleting row k and column m.

Theorem 3: Write C(Y) as

$$C(Y) = P_o + \alpha P_\alpha \qquad (4)$$

where P_o is the sum of those terms of C(Y) which do not contain α, and αP_α is the sum of those terms which contain α (this form is always possible since the network is assumed to be linear).

Then $P_o = C[Y|_{\alpha=0}]$

and $P_\alpha = (-1)^{j+m}C(Y_\alpha)$

where Y_α is a matrix obtained by modifying Y as follows:

(1) add row j to row i.
(2) add column m to column ℓ.
(3) delete row j and column m.

Proof: By definition P_o is the sum of those terms of C(Y) which do not contain the parameter α and hence is given by

$$P_o = C[Y|_{\alpha=0}]$$

To find P_α we begin by noting that as a result of Theorem 1

$$C(Y) = (-1)^{j+m}\det(Y_{jm})$$

Now the matrix Y_{jm} has only one position that contains the parameter α, namely (i,ℓ). Thus, P_α is simply the cofactor of position (i,ℓ) of Y_{jm}.

$$P_\alpha = (-1)^{j+m}(-1)^{i+\ell}\det[(Y_{jm})_{i\ell}]$$

Because only rows i and j and columns ℓ and m are modified or deleted in forming Y_α, it follows that

$$\det[(Y_{jm})_{i\ell}] = (-1)^{i+\ell}C_{i\ell}(Y_\alpha)$$

But noting that Y_α is also an indefinite admittance matrix, this becomes

$$\det\left[(Y_{jm})_{i\ell}\right] = (-1)^{i+\ell}C(Y_\alpha)$$

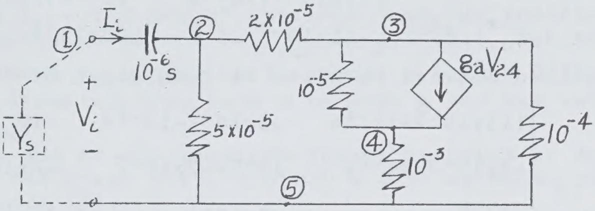
Thus,

$$P_\alpha = (-1)^{j+m}C(Y_\alpha) \qquad \text{QED}$$

An example is now given to illustrate how a network function can be derived using a "parameter extraction process" based on Eq. (4).

Example 1: Find Z_i for the network in Fig. 2, keeping g_a in symbolic form.

Figure 2



The admittance Y_s has been included in Fig. 2 as specified by Theorem 2. From Theorem 2 it follows that

$$C(Y) = Y_sP_{Y_s} + P_o$$

such that

$$Z_i = \frac{V_i}{I_i} = \frac{P_{Y_s}}{P_o}$$

The indefinite admittance matrix Y is given by

$$Y = \begin{bmatrix} Y_s+10^{-6}s & -10^{-6}s & 0 & 0 & -Y_s \\ -10^{-6}s & 10^{-6}s+7\times10^{-5} & -2\times10^{-5} & 0 & -5\times10^{-5} \\ 0 & g_a-2\times10^{-5} & 13\times10^{-5} & -10^{-5}-g_a & -10^{-4} \\ 0 & -g_a & -10^{-5} & 101\times10^{-5}+g_a & -10^{-3} \\ -Y_s & -5\times10^{-5} & -10^{-4} & -10^{-3} & 115\times10^{-5}+Y_s \end{bmatrix}$$

Let Y_s be the first parameter to extract from C(Y). In this case $i=\ell=1$, $j=m=5$ so that

$$C(Y) = C\left[Y\middle|_{Y_s=0}\right] + Y_sC\left[Y_{Y_s}\right]$$

or

$$C(Y)=C\left[Y\middle|_{Y_s=0}\right]+Y_sC\begin{bmatrix} 115\times10^{-5}+10^{-6}s & -5\times10^{-5}-10^{-6}s & -10^{-4} & -10^{-3} \\ -5\times10^{-5}-10^{-6}s & 10^{-6}s+7\times10^{-5} & -2\times10^{-5} & 0 \\ -10^{-4} & g_a-2\times10^{-5} & 13\times10^{-5} & -10^{-5}-g_a \\ -10^{-3} & -g_a & -10^{-5} & 101\times10^{-5}+g_a \end{bmatrix}$$

Now the parameter g_a must be extracted from $C\left[Y\big|_{Y_s=0}\right]$ and $C(Y_{Y_s})$ where in both cases $i=3$, $\ell=2$, $j=m=4$.

$$C(Y) = C\left[(Y\big|_{Y_s=0})\big|_{g_a=0}\right] + g_a C \begin{bmatrix} 10^{-6}s & -10^{-6}s & 0 & 0 \\ -10^{-6}s & 10^{-6}s+7\times10^{-5} & -2\times10^{-5} & -5\times10^{-5} \\ 0 & 98\times10^{-5} & 12\times10^{-5} & -11\times10^{-4} \\ 0 & -105\times10^{-5} & -10^{-4} & 115\times10^{-5} \end{bmatrix}$$

$$+ Y_s C\left[Y_{Y_s}\big|_{g_a=0}\right] + Y_s g_a C \begin{bmatrix} 115\times10^{-5}+10^{-6}s & -105\times10^{-5}-10^{-6}s & -10^{-4} \\ -5\times10^{-5}-10^{-6}s & 10^{-6}s+7\times10^{-5} & -2\times10^{-5} \\ -11\times10^{-4} & 98\times10^{-5} & 12\times10^{-5} \end{bmatrix}$$

The desired network function is then given by

$$Z_i = \frac{C\begin{bmatrix} 115\times10^{-5}+10^{-6}s & -5\times10^{-5}-10^{-6}s & -10^{-4} & -10^{-3} \\ -5\times10^{-5}-10^{-6}s & 10^{-6}s+7\times10^{-5} & -2\times10^{-5} & 0 \\ -10^{-4} & -2\times10^{-5} & 13\times10^{-5} & -10^{-5} \\ -10^{-3} & 0 & -10^{-5} & 101\times10^{-5} \end{bmatrix} + g_a C\begin{bmatrix} 115\times10^{-5}+10^{-6}s & -105\times10^{-5}-10^{-6}s & -10^{-4} \\ -5\times10^{-5}-10^{-6}s & 10^{-6}+7\times10^{-5} & -2\times10^{-5} \\ -11\times10^{-4} & 98\times10^{-5} & 12\times10^{-5} \end{bmatrix}}{C\begin{bmatrix} 10^{-6}s & -10^{-6}s & 0 & 0 & 0 \\ -10^{-6}s & 10^{-6}s+7\times10^{-5} & -2\times10^{-5} & 0 & -5\times10^{-5} \\ 0 & -2\times10^{-5} & 13\times10^{-5} & -10^{-5} & -10^{-4} \\ 0 & 0 & -10^{-5} & 101\times10^{-5} & -10^{-3} \\ 0 & -5\times10^{-5} & -10^{-4} & -10^{-3} & 115\times10^{-5} \end{bmatrix} + g_a C\begin{bmatrix} 10^{-6}s & -10^{-6}s & 0 & 0 \\ -10^{-6}s & 10^{-6}s+7\times10^{-5} & -2\times10^{-5} & -5\times10^{-5} \\ 0 & 98\times10^{-5} & 12\times10^{-5} & -11\times10^{-4} \\ 0 & -105\times10^{-5} & -10^{-4} & 115\times10^{-5} \end{bmatrix}}$$

If Z_i is to be evaluated at a single frequency, say s=j2, then the cofactors have only complex numerical entries and can be evaluated using standard numerical methods. However, if s is to be kept as a variable then numerical algorithms such as Gauss elimination are not applicable. One possible technique would be to apply the state-space approach to the networks corresponding to the above indefinite admittance matrices.

A "PARAMETER EXTRACTION PROCESS" BASED ON GRAPH MODIFICATION

Definitions:

G(Y) = the directed graph* obtained from the indefinite admittance matrix $Y=\{y_{pq}\}$ as follows: For each element y_{pq}, $p\neq q$, in Y, create a branch G(Y) which is directed from node p to node q and has weight $-y_{pq}$.

$T[G(Y)]$ = the sum of the tree admittance products of the directed graph G(Y); $T[G(Y)] = C(Y)$ as shown in reference [2].

X^{ij} = the matrix formed from X by interchanging rows i and j.

It is the purpose of this section to reformulate the matrix modification method into a graph modification technique. Although Theorem 3 can be interpreted directly in terms of modifying graphs, a somewhat more efficient scheme can be formulated from the following theorem.

Theorem 4: $C(Y^{ij}) = -C(Y) \qquad i\neq j \qquad (5)$

Proof: The theorem is obviously true if n=2. For the case n=3 we have only to prove

$$C_{kl}(Y^{ij}) = -C_{kl}(Y) \qquad k\neq i\neq j$$

because Y^{ij} is also an indefinite admittance matrix and from Theorem 1 all the cofactors of Y^{ij} are equal. An examination of $C_{kl}(Y^{ij})$ and $C_{kl}(Y)$ immediately shows that the two determinants differ only in an interchange of two rows. Therefore, (5) is valid. QED

Again let a general network element associated with a parameter α be represented in the indefinite admittance matrix Y by (3). Observe that the element can be classified as one of the following types depending on the position of its entries in Y.

Type (A): Passive if $i=\ell$ and $j=m$.
Type (B): Active 3-terminal if $i\neq\ell$ and $j=m$.
Type (C): Active 3-terminal if $i=\ell$ and $j\neq m$.
Type (D): Active 4-terminal if $i\neq\ell$ and $j\neq m$.

As a result of Theorem 4 an element can be made to be passive (Type (A)) without changing the value of the cofactors (except for a possible change in sign) by interchanging certain rows of Y as follows:

---

* This graph differs from that used by Chen [3] in that our graph is derived from the indefinite admittance matrix and hence is independent of a reference node.

(a) if α corresponds to type (B) in G(Y), it will correspond to type (A) in $G(Y^{i\ell})$ where $C(Y) = -C(Y^{i\ell})$.

(b) if α corresponds to type (C) in G(Y), it will correspond to type (A) in $G(Y^{jm})$ where $C(Y) = -C(Y^{jm})$.

(c) if α corresponds to type (D) in G(Y), it will correspond to type (A) in $G\left[(Y^{i\ell})^{jm}\right]$ where $C(Y) = C\left[(Y^{i}\ )^{jm}\right]$.

From (a), (b), and (c) it follows that

$$T\left[G(Y)\right] = K\ T(G_{pas.\alpha}) \tag{6}$$

where

$$G_{pas.\alpha} = \begin{cases} G(Y) & \text{if } \alpha \text{ is type (A) in } G(Y) \\ G(Y^{i\ell}) & \text{if } \alpha \text{ is type (B) in } G(Y) \\ G(Y^{jm}) & \text{if } \alpha \text{ is type (C) in } G(Y) \\ G\left[(Y^{i\ell})^{jm}\right] & \text{if } \alpha \text{ is type (D) in } G(Y) \end{cases}$$

and

$$K = \begin{cases} +1 & \text{if } \alpha \text{ is type (A) or type (D) in } G(Y) \\ -1 & \text{if } \alpha \text{ is type (B) or type (C) in } G(Y) \end{cases}$$

Thus, we have reduced the problem of extracting an active type parameter α from G(Y) to that of extracting a passive type parameter α from a modified graph, $G_{pas.\alpha}$. But extracting a passive type α from a graph is a simple extension of a well known technique, [4] and [5], that can be summarized by the equation

$$T(G_{pas.\alpha}) = T(G_{pas.\alpha}|_{\alpha=0}) + \alpha T\left[(G_{pas.\alpha})_{\alpha}\right] \tag{7}$$

where $G_{pas.\alpha}|_{\alpha=0}$ is formed from $G_{pas.\alpha}$ by eliminating the directed branches associated with α and $(G_{pas.\alpha})_{\alpha}$ is formed from $G_{pas.\alpha}$ by coalescing the two terminal nodes of the directed branches associated with α.

It remains to show how $G_{pas.\alpha}$ can be obtained from G(Y) directly without resorting to the indefinite admittance matrix. Define "node interchange" between nodes i and ℓ of the graph G(Y) as the graph modification necessary in formulating $G(Y^{i\ell})$ from the graph G(Y), a process which can be summarized as follows (refer to example at end of this section):

Initially let $G(Y^{i\ell})$ be a graph of n nodes and no branches.

(I) If a branch is directed from node i to node N $(N=1,2,\ldots,n;\ N\neq\ell)$ in G(Y), create a branch in $G(Y^{i\ell})$ directed from node ℓ to node N, $N=1,2,\ldots,n$.

(II) Use the negative of the sum of admittance weights of each branch directed out of node i in G(Y) as the weight of a branch directed from node ℓ to node i in $G(Y^{i\ell})$.

(III) Repeat (I) and (II) after interchanging i and ℓ.

(IV) For each branch of G(Y) not involved in the manipulations in (I), (II), or (III), create an identical branch in $G(Y^{i\ell})$.

Label the graph configuration of a general active element (defined by (3)) as shown in Fig. 3(a).

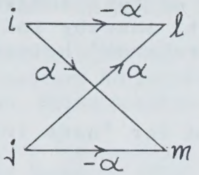


Figure 3(a): Graph for an Active Element α

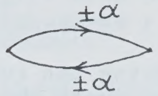


Figure 3(b): Graph for a Passive Element α

Consider the set of node pairs of the graph in Fig. 3(a) such that when each node pair in the set is coalesced, the graph in Fig. 3(a) will be reduced to a graph of a passive element as shown in Fig. 3(b). $G_{pas.\alpha}$ can be obtained from the graph G(Y) by performing a "node interchange" on each node pair in this set.

Example 2: Using graph modification techniques, again find Z_i for the network in Fig. 2 keeping g_a in symbolic form. Starting with G(Y) given in Fig. 4 and choosing Y_s as the first parameter to extract, Eq. 7 gives

$$T[G(Y)] = T[G|_{Y_s=0}] + Y_s T[G_{Y_s}]$$

where $G|_{Y_s=0}$ and G_{Y_s} are given in Figs. 5 and 6.

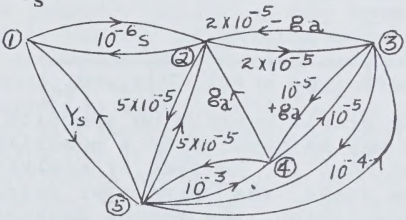

①
10⁻⁶s
2×10⁻⁵ - ga
③
②
2×10⁻⁵
5×10⁻⁵
ga
10⁻⁵ +ga
Ys
10⁻⁵
5×10⁻⁵
④
10⁻⁴
10⁻³
⑤

Figure 4: G(Y)

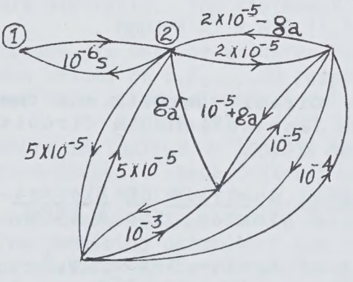

①
②
2×10⁻⁵ - ga
10⁻⁶s
2×10⁻⁵
ga
10⁻⁵ +ga
5×10⁻⁵
5×10⁻⁵
10⁻⁵
10⁻⁴
10⁻³

Figure 5: $G|_{Y_s=0}$

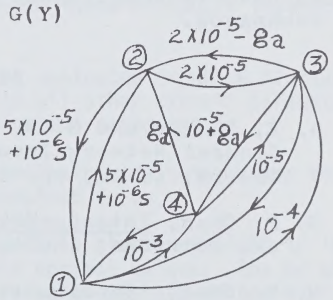
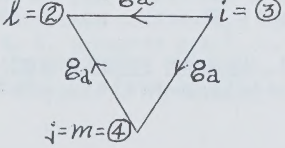

2×10⁻⁵ - ga
②
2×10⁻⁵
③
5×10⁻⁵ +10⁻⁶ s
ga
10⁻⁵ +ga
5×10⁻⁵ +10⁻⁶s
④
10⁻⁵
10⁻⁴
10⁻³
①

Figure 6: G_{Y_s}

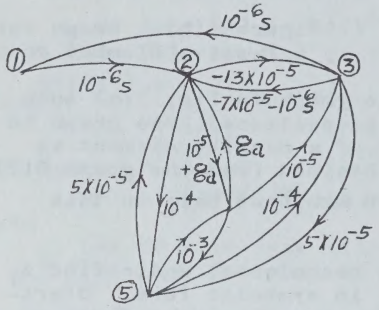
The parameter g_a must now be extracted from each of the above graphs. In both graphs i=3, =2, j=m=4 as shown in Fig. 7.

Figure 7


l=②
-ga
i=③
ga
ga
j=m=④

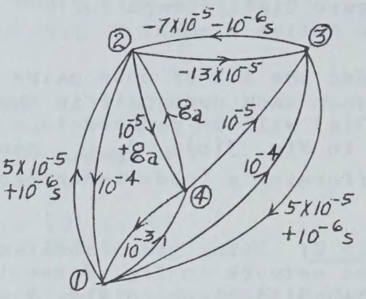
Now noting that coalescing nodes 2 and 3 of this subgraph will reduce it to a graph of a passive element, thereby indicating that in both $G|_{Y_s=0}$ and G_{Y_s} a "node interchange" between nodes 2 and 3 is necessary.

Using the rules previously developed for "node interchanging" produces the following graphs.



$10^{-6}s$
①
②
③
$10^{-6}s$
-13×10^{-5}
$-7\times10^{-5}-10^{-6}s$
$10^{-5}+g_a$
g_a
10^{-5}
5×10^{-5}
10^{-4}
10^{-4}
10^{-3}
5×10^{-5}
⑤

Figure 8: $(G|_{Y_s=0})_{pas.g_a}$



②
$-7\times10^{-5}-10^{-6}s$
③
-13×10^{-5}
$10^{-5}+g_a$
g_a
10^{-5}
$5\times10^{-5}+10^{-6}s$
10^{-4}
10^{-4}
④
$5\times10^{-5}+10^{-6}s$
10^{-3}
①

Figure 9: $(G_{Y_s})_{pas.g_a}$

The parameter g_a now corresponds to a passive element in the graphs of Fig. 8 and Fig. 9 and can easily be extracted. Then

$$Z_i = \frac{-T\left[G_{Y_s}|_{g_a=0}\right] - g_a T\left[(G_{Y_s})_{g_a}\right]}{-T\left[(G|_{Y_s=0})|_{g_a=0}\right] - g_a T\left[(G|_{Y_s=0})_{g_a}\right]}$$

The four resulting graphs contain no symbolic parameters (other than s) and hence may be handled by a numerical type program. It should be noted that if, instead of applying the numerical subroutine at some stage, the "parameter extraction process" is continued to completion, the method becomes a tree enumeration technique.

REFERENCES

(1) G. E. Sharpe and G. P. H. Styan, "Circuit Duality and the General Network Inverse," IEEE Transactions on Circuit Theory, CT-12, pp. 22-27; March 1965.

(2) S. P. Chan, Introductory Topological Analysis of Electrical Networks, (Holt, Rinehart, & Winston, Inc.) 1969.

(3) W. K. Chen, "Topological Analysis of Active Networks," IEEE Transactions on Circuit Theory, March 1965.

(4) W. Feussner, "Uber Stromverzweigung in Netzformigen Leitern," Ann. Phys., vol. 9, pp. 1304-1329; 1902.

(5) L. Weinberg, "Kirchhoff's 'Third and Fourth Laws,'" IRE Transactions on Circuit Theory, pp. 8-30; March 1958.

(6) S. Seshu and M. B. Reed, Linear Graphs and Electrical Networks, (London: Addison-Wesley) 1961.

SYMBOLIC NASAP FOR LINEAR CIRCUIT DESIGN

RAYMOND S. SCHWARTZ
Northeastern University & Alford Manufacturing Company
Boston, Massachusetts — Winchester, Massachusetts

Recent improvements in the loop finding-coding-sorting algorithm employed by Symbolic NASAP in the application of Mason's formula to transfer function evaluation [1] now make entirely symbolic solutions a practical reality. The algorithm specifies use of integer arithmetic to label loop sets resulting in greater efficiency and improved accuracy. As each loop set is found it is assigned a distinct integer label from which the original loop set may be regenerated. The magnitude of the label indicates which network branches are contained in the loop set; the sign indicates the overall loop-set sign. The labels are sorted into descending order of magnitude for more efficient decoding and stored in a loop-occurrence array using sequential memory locations to minimize storage requirements. If two labels of the same magnitude but opposite sign are found, both are eliminated thereby suppressing generation of spurious transfer function coefficients. The completed loop-occurrence array contains a coded symbolic solution. This solution is then evaluated for the nominal parameter values yielding polynomials in the complex frequency parameter s.

Transfer functions may be evaluated either symbolically in terms of the element names or numerically in terms of the element values. However, the symbolic solution generally contains so many terms that its direct application to design is limited. Instead, parameter modification proves more useful. Since each new solution requires merely the evaluation of a stored symbolic solution, this procedure is quite economical. Figure 1 compares the time required by Symbolic NASAP and the original NASAP to find several solutions to a two-stage transistor amplifier model.

The MODIFY instruction is a most powerful tool in design applications. Parameter modification may be used for on-line design or for automatic design in conjunction with iterative optimization or Monte Carlo analysis overlays. In Symbolic NASAP several forms of MODIFY instructions are available. The statement

```
MODIFY R1 = 5,8,11,14,17,20     ,
```

causes six sets of transfer functions to be evaluated in which R1 takes the values of 5,8,...,20 respectively while all other circuit parameters are held at their nominal values. An equivalent statement is

```
MODIFY (R1 = 5,20,3)     ,
```

where an implied do loop is indicated by enclosing the instruction in parentheses. These instructions may be combined as in

```
MODIFY C1 = .5,  (R2 = 10,1000,10)     ,
```

which results in C1 being held at .5 while R2 is stepped from 10 to 1000. The remaining parameters are fixed at their nominal values. Any or all parameters may be changed since each modification results in the evaluation of a new solution. Figure 2 illustrates the familiar Wien-bridge circuit, its coding for solution by NASAP, and its symbolic solution. Figure 3 illustrates use of the MODIFY instruction to obtain data for a root locus plot. The parameter being varied is the gain of the operational amplifier.

References: [1] R. E. Bach, Jr. and P. A. Olendzenski, NASAP FOR ON-LINE APPLICATIONS, Northeastern University, Boston, Massachusetts, January, 1969

[2] R. S. Schwartz and R. E. Bach, Jr., A SYMBOLIC SOLUTION ALGORITHM FOR DESIGN OF LINEAR NETWORKS, Thirteenth Midwest Symposium on Circuit Theory, Minneapolis, Minnesota, May, 1970

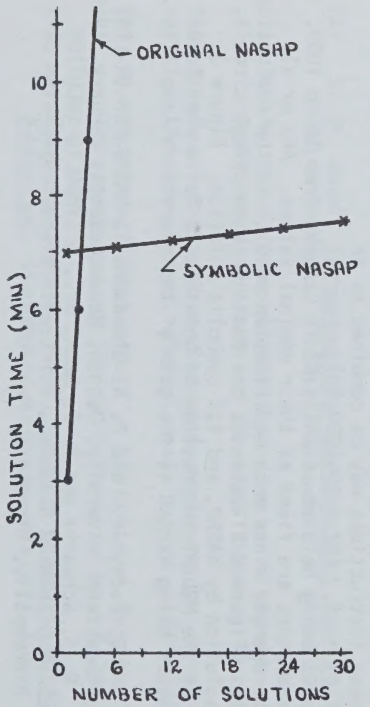
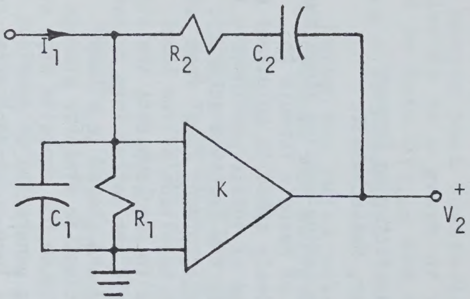


Fig. 1. Comparison of the time required to evaluate several solutions for a two-stage transistor amplifier model.

(a) Wien-bridge circuit:



(b) Coding and symbolic solution:

```
NASAP PROBLEM
* WIEN-BRIDGE CIRCUIT
I1 1 2 1
R1 2 1 1
C1 2 1 1
R2 2 3 1
C2 3 4 1
V2 1 4 1 VR1
OUTPUT
VV2/II1
ROOTS
VV2/II1
EXECUTE
SYMBOLIC SOLUTION


NUMERATOR POLYNOMIAL OF VV02/II01

 LOC   LABEL   LOOP PRODUCT
  1     -27    (R01V02 / C02R02) S ** -1
  2      -3    (R01V02) S **  0


DENOMINATOR POLYNOMIAL

 LOC   LABEL   LOOP PRODUCT
               ( 1 ) S **  0
  3      30    (C01R01 / C02R02) S **  0
  4      24    (  1 / C02R02) S ** -1
  5     -11   -(R01V02 / R02) S **  0
  6      10    (R01 / R02) S **  0
  7       6    (C01R01) S **  1
```

(c) Transfer function: $$(V_2/I_1) = \frac{KR_1(1 + R_2C_2S)}{1 + [R_1C_1 + R_2C_2 + R_1C_2(1 - K)]S + R_1C_1R_2C_2S^2}$$

Fig. 2. Wien-bridge circuit with NASAP symbolic solution.

(a) Root-locus data:

```
MODIFY ( V2=0,10,1 )


MODIFYING V2 TO  0


K =  0

NUMERATOR ROOTS (ZEROS) OF VV02/II01
   S = -1.0000E 00


DENOMINATOR ROOTS (POLES)
   S = -3.8197E-01
   S = -2.6180E 00
           |
           |
           |
MODIFYING V02 TO  3.0000E 00


K =  3.0000E 00

NUMERATOR ROOTS (ZEROS) OF VV02/II01
   S = -1.0000E 00


DENOMINATOR ROOTS (POLES)
   S =            0   + J 1.0000E 00
   S =            0   - J 1.0000E 00
```

(b) Root locus:

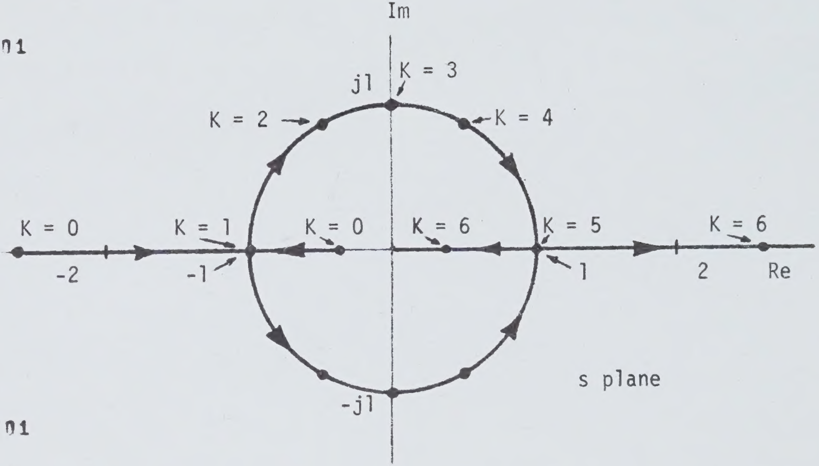

Im
j1
K = 3
K = 2
K = 4
K = 0
K = 1
K = 0
K = 6
K = 5
K = 6
-2
-1
1
2
Re
s plane
-j1

Fig. 3. Illustrating use of MODIFY statement to obtain root-locus data for the Wien-bridge circuit.

DIGITAL COMPUTER
LABORATORY
LIBRARY

LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

UNIVERSITY OF ILLINOIS-URBANA
621.31921AL54P C003
PROCEEDINGS.$URBANA ILL
8 1970
3 0112 006279654